

UNIVERSITY OF PADUA, DEPARTMENT OF GENERAL PSYCHOLOGY PhD THESIS IN BRAIN MIND
AND COMPUTER SCIENCE, XXXVIII SERIES

APICURON: a database for credit & Recognition of biocuration activities

Thesis written with the financial contribution of the European Union - Next Generation EU -
National Recovery and Resilience Plan (PNRR), M4C2 Investment 3.1.
Strengthening of Research Infrastructures. "Elixir x Next Generation IT: Consolidamento
dell'Infrastruttura Italiana per i Dati Omici e la Bioinformatica - NextGenIT" Program - CUP
B53C22001800006

Kamel Eddine Adel Bouhraoua

ID 0023000596

SUPERVISOR

Prof. Silvio Tosatto

University of Padua

INTERNAL SUPERVISOR

Dr. Federica Quaglia

University of Padua

CO-SUPERVISOR

Dr. Luciano Gamberini

University of Padua

ACADEMIC YEAR 2025/2026

I would like to begin by expressing my profound gratitude to my parents, whose unwavering belief in me has been essential to my journey. Their countless sacrifices have ensured that I was provided with the best conditions for my growth and development.

To my family and friends, your continued support and faith in my capabilities have been invaluable, and I am forever thankful for your presence in my life. A special word of thanks goes to my esteemed supervisors, Silvio Tosatto and Federica Quaglia. Their relentless guidance, support, expertise, and encouragement have been instrumental every step of the way. I cannot express enough how grateful I am for everything they have done to bring out the best in me. I also want to acknowledge my colleagues, with whom I've had the privilege of spending the last three years. Together, we've shared numerous memorable moments that I will cherish forever.

I'm deeply grateful to the jury members for their time, patience, and willingness to assess and examine my work.

Lastly, my heartfelt appreciation goes out to every individual, near or far, who offered support in any form, even if it was just a warm smile. Your gestures, no matter how big or small, made a world of difference.

Thank you all!

Contents

Abstract	ix
Introduction	1
0.1 Biocuration	2
0.2 Biocuration in a global context	2
0.3 The Reward Discrepancy	2
0.4 APICURON: The Initial Solution	3
0.5 Summary of Contributions	4
I Credit & Recognition	7
1 Credit & Recognition	9
1.1 Shift Towards Data-centric Science	9
1.2 The Acknowledgement Pathway	10
1.3 Global Frameworks for Research Assessment Reform	11
1.3.a The Declaration On Research Assessment (DORA)	11
1.3.b Principles for Responsible Metrics and Integrity	12
1.3.c The Coalition for Advancing Research Assessment (CoARA)	12
1.3.d Addressing the Invisible Labor: The Hidden REF	13
1.3.e The Implementation Gap	13
1.4 Standards and Taxonomies	14
1.4.a Contributor Roles and Taxonomies CRediT	14
1.4.b Software Citation Standards	15
1.4.c Provenance and granularity gap	15
1.4.d The need for an Aggregation Framework	16
1.4.e Towards a Federated Ecosystem for Recognition	16
1.4.f The Architectural Components	17
1.4.g Strategic Placement of APICURON	18
II APICURON	19
2 APICURON	21

2.1	The Recognition Framework	22
2.1.a	Metrics, Engagement, and Incentive Mechanisms	22
	Temporal Segmentation	23
	Comparative Context	23
	Badges	23
	Medals	23
	Contextual Fairness via Contribution Groups	24
2.2	Technical Implementation	25
2.2.a	Database Structure	26
2.2.b	Authentication and Access Control	30
	Identity Management and Provenance	30
	Delegated Authentication via ORCID	30
	Role-Based Access Control	30
2.2.c	Data Ingestion	31
	The Timely Submission	31
	The Bulk Submission	31
2.2.d	The Event-Driven Processing Engine	33
	Unified Event Abstraction	33
2.2.e	Data Distribution & Dissemination	39
	The granularity mismatch	39
III	DisProt	41
3	DisProt	43
3.1	Operational Context & Legacy constraints	43
3.1.a	Semantic Integration Challenges	44
3.1.b	Decoupling Execution from Attribution	44
3.1.c	Legacy Modernization Strategy	44
3.2	The Deposition Process	48
3.3	Operational Recognition	49
3.4	Impact Analysis	50
IV	The APICURON Implementation Study	51
4	The APICURON Implementation Study	53
4.1	Objectives and scope	54
4.2	Methodology	54
4.2.a	Implementation	55
4.2.b	Concepts	55
4.2.c	Quantitative analysis	57
4.2.d	Qualitative Analysis	57
4.3	Results	58
4.3.a	Qualitative Analysis	58

4.3.b	Quantitative Analysis	58
4.3.c	DisProt	64
4.3.d	Reactome	67
4.3.e	Pfam	71
4.3.f	IntAct	74
4.3.g	Complex Portal	75
4.3.h	BioModels	77
4.4	Implementation & Analysis Outcomes	79
4.4.a	Reports on Biocuration Activities	79
4.4.b	Validated Data Model	79
4.4.c	Sociotechnical Findings	79
4.4.d	The Consent Gap	80
4.4.e	Broadening scope	80
4.5	Conclusion	81
V	The Consent Management System	83
5	The Consent Management System	85
5.1	Context	85
5.2	Regulatory Basis	86
5.2.a	Definitions	86
5.2.b	Lawfulness of processing	86
5.2.c	Demonstration of Compliance	87
5.2.d	Transparency	87
5.3	System Overview	87
5.3.a	Technical Design	88
5.3.b	Architecture: Components & Tools	89
	Web component	89
5.3.c	The web server	89
5.3.d	The Database	89
5.3.e	Workflows	90
	Gather Consent Data	90
	Query Consent Data	91
VI	Credit & Recognition for Software	93
6	Credit & Recognition for Software	95
6.1	The Software Attribution Paradox	96
6.1.a	Process over Product Recognition	96
6.1.b	The Commit Fallacy	97
6.1.c	The Semantic Gap	97
6.2	The Research Software Credit Schema	97
6.2.a	Context-Aware Categorization	99

6.3	Technical Implementation	100
6.3.a	The Data Capture Challenge	100
6.3.b	The Proof of concept	103
6.4	The Logic Layer: Stateful Contribution Mapping	103
6.4.a	The Shift from Stateless to Stateful Recognition	103
6.5	The Acknowledgement layer	105
6.5.a	Ecosystem Integration and Feedback Loops	105
VII	Conclusions	107
	Conclusions	109
6.6	Implications for the Field	110
6.7	Perspectives	110
6.7.a	From Gamification to Career Utility	110
6.7.b	Semantic Middleware in a Federated Ecosystem	110
6.7.c	Closing the Compliance and Trust Gap	111
6.7.d	The Human role in Data Production	111
	Bibliography	115
	Acknowledgments	121

Abstract

The consolidation of massive data sources over the past decade has revolutionized scientific inquiry, placing data at the center of modern research paradigms[18]. This reliance has led to a paradigm where all decision-making is data-driven. This shift is equally transformative in scientific research, particularly in biology, where recent, significant breakthroughs have been directly attributable to data-driven approaches leveraging Artificial Intelligence[17].

In this context, the demand for Findable, Accessible, Interoperable, and Reusable (FAIR) and high-quality data is paramount for sustainable, modern research. This necessity highlights the critical role of Biocuration: the essential process of manually transforming unstructured biological information into a structured, computable, and FAIR format. However, accurately quantifying, tracking, and formally attributing the intense intellectual effort of biocurators remains a significant challenge.

The work described in this PhD Thesis addresses this gap through the development and implementation of APICURON, a web-based service designed to provide formalized Credit & Recognition for Biocuration efforts. The system aggregates curation events from disparate resources, utilizing a standardized schema to capture curator activity. With a granular efficient computation workflow that allows a continuous timely reaction to user activity.

Additionally, to demonstrate APICURON's efficacy and interoperability, this work also encompassed the development and major update of DisProt, the database of Intrinsically Disordered Proteins. DisProt served as the primary pilot implementation for APICURON. By operating simultaneously as the infrastructure developer and the database maintainer, this work validated APICURON's ability to capture granular curation activity and display attribution metrics within a live, production-level resource.

Crucially, the system's maturity was accelerated by a group of early adopters, specifically through integration with major EMBL-EBI resources. While technically successful, this expansion revealed significant privacy-related challenges regarding General Data Protection Regulation (GDPR) compliance. The regulatory requirement to exclude data lacking explicit user consent introduced workflow complexities and underscored the need for a standardized privacy mechanism. In response, this thesis proposes the APICURON Consent Management Platform: an open solution specifically designed to manage consent in Data Sharing scenarios, a critical use case often overlooked by commercial offerings, thereby ensuring lawful data processing and enabling seamless future integrations.

Building on this successful implementation along with its challenges, the scope of the thesis extends to address a broader challenge: the lack of recognition for diverse research outputs beyond traditional publications. Modern researchers frequently engage in software development, dataset creation, and educational training, activities often invisible in scholarly publications and standard indicators. Leveraging its flexible architecture, APICURON was adapted to track these varied workflows.

Ultimately, this thesis establishes a comprehensive ecosystem for scientific attribution beyond traditional publications that bridges technical implementation with regulatory compliance. By making hidden labor visible and overcoming privacy barriers, this work lays the foundation for a more equitable and sustainable reward system suited to the era of data-driven research.



Introduction

The central aim of this doctoral thesis is to address the significant challenge associated with the accurate quantification, tracking, and crediting of biocuration activities. This work details the enhancements made to APICURON to surmount these challenges. Specifically, it employs the DisProt use-case to demonstrate how APICURON facilitates the attribution of credit within complex curation workflows. Furthermore, overcoming these challenges permits APICURON to establish a versatile system applicable to diverse use-cases beyond the scope of biocuration.

To contextualize this work effectively, a formal landscape analysis must first be established to define the current state of the art, particularly regarding how researchers are formally recognized for their increasingly sparse efforts and the consequential impact of this recognition on their professional careers.

The consolidation of massive data sources over the past decades alongside advancements in Machine learning and Artificial Intelligence has placed data in the center of scientific inquiry. This transition has engendered a paradigm shift, introducing a new form of empiricism often described as “the end of theory”.[18]. Consequently, decision-making has become increasingly data-driven, a trend that is particularly transformative in biology where recent breakthroughs are directly attributable to Artificial Intelligence.

With the advent of Large Language Models (LLMs), scholars now recognize the necessity of establishing a robust “chain of production” for software technology, driven by the industrialization of Artificial Intelligence[40]. While this chain relies heavily on modern cloud infrastructure and vast data repositories, one of its most critical yet overlooked components is the human investment. The efficacy of modern AI is not solely a product of algorithms, but of the collective contribution of structured data generated by human experts[36].

0.1 Biocuration

Biocuration is the process of extracting, interpreting, and organizing biological information from the scientific literature and other sources into a structured, machine-readable format, typically within biological databases. Biocurators are typically PhD-level biologists often with lab bench experience[5]. The task involves synthesizing and integrating data from various sources, such as peer-reviewed publications, major initiatives like the Encyclopedia of DNA Elements[1], and conference abstracts. To accurately document a researcher's conclusion and the supporting evidence, they must directly contact authors for clarifications, analyze supplementary material, and resolve identifiers. Effective biocuration relies on the ability to distill the current "best view" from potentially conflicting sources.

0.2 Biocuration in a global context

After the well established scalability of expert curation in one of the largest protein sequence databases in the world[29]. It is evident that Biology in the era of data driven science relies heavily on the modern data infrastructure as well as human investment to produce reliable quality data. This places biocuration as a fundamentally valuable intermediary between raw scientific discovery and its computational utility. Without meticulous biocuration, the vast and ever-growing corpus of biological literature remains inaccessible to computational analysis, data integration, and knowledge discovery tools. However, biocuration is an intensely time-consuming and cognitively demanding process. It requires specialized biological domain expertise, careful reading and interpretation of primary research articles, adherence to complex and evolving curation standards (ontologies, controlled vocabularies), and the manual entry and verification of data.

0.3 The Reward Discrepancy

Despite the critical role of biocuration, it remains largely invisible in terms of formal academic reward. This discrepancy is not only a matter of overlooked effort, but of a fundamental structural incompatibility between the nature of biocuration and the traditional indicators of academic success.

Traditional Scholarly recognition is typically designed to reward the end result of science, novel discoveries encapsulated in narrative-driven publications. In contrast, biocuration is part of a larger continuous chain of production, similar

to industrial infrastructure maintenance rather than discrete discovery. Consequently, the maintenance of high-quality data, no matter how rigorous, cannot effectively compete with the novelty of a traditional research paper. The current academic reward system is therefore ill-equipped to quantify this essential chain of production, creating a gap where the creators of critical research infrastructure are excluded from professional advancement.

Moreover, several sparse efforts have been made to attribute proper credit to the work of biocurators in various manually curated biological data resources[14]. A notable example is the Pfam[13] database of protein families, which leveraged the Open Researcher and Contributor ID. Within this framework, biocurators can link their contributions directly to their Open Researcher and Contributor ID (ORCID) profiles, effectively claiming authorship of specific Pfam-curated entries. However, while valuable, these solutions remain fragmented and inherently resource-specific. A critical constraint is that the unit of contribution is often selected as the final static object rather than the specific actions taken to produce it. By recording only the end result, this approach inherently simplifies the curation workflow, resulting in ambiguous forms of credit such as authorship. These vague forms of credit fail to differentiate between a complex conceptual validation and a minor annotation update. Furthermore, this coarse granularity fails to capture the intricacies of the curation workflow; all entries are credited equally. Consequently, these fragmented solutions lack a fair representation of effort capable of reflecting the depth of the scientific effort undergone in the biocuration workflow.

0.4 APICURON: The Initial Solution

This PhD thesis directly addresses the lack of credit & recognition for biocuration efforts and research activities beyond traditional publications in the scientific landscape. To remedy this issue, this research focuses on the expansion and development of APICURON.

APICURON is a web-based service specifically engineered to provide transparent, formalized, and standardized Credit & Recognition for all documented Biocuration efforts across diverse resources.

APICURON leverages the very basic, three-axiom, schema based on simple PROV-O[30, 38] entities and properties to represent a single activity occurrence. This basic representation allows APICURON to operate on simple building blocks to achieve Credit & Recognition for biocuration activities.

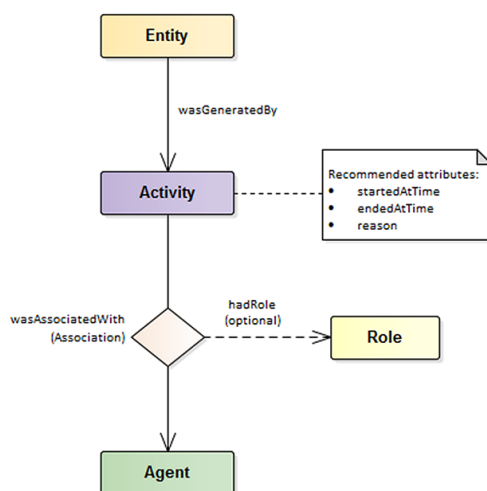


Figure 1: Basic schema linking Entities to Activities and Activities to Agents. Roles are optional from the RDA/TDWG Working Group, source: [38]

The primary benefit of APICURON’s data representation is its focus on the activity scope rather than the data scope. Instead of merely processing the static entity, which often reveals little about the developmental effort, APICURON processes the actions and activities associated with it. This model allows APICURON to represent an entity not as a simple object, but as a comprehensive series of context relevant events putting the human effort at the center of the data model.

0.5 Summary of Contributions

Prior to the work outlined in this thesis, the initial version of APICURON focused on crediting biocuration contributions. Within this scope, it relied on an internal pipeline that run periodically on contribution data submitted by the integrated partners. This design allowed contributions to be recognized only after each periodic execution of the pipeline. With its initial scope, APICURON had a limited set of integrated platforms, the onboarding process was slow and required relevant manual intervention, restricting the initial set of resources to only 3 partners.

To address the limitations of the initial system, the specific contributions outlined in this thesis are as follows:

- **Architectural Enhancement:** he implementation of an event-based system to process contributions at a granular level, leveraging the activity scope

offered by the data model and enabling on-the-fly processing to reflect the impact of contributions as efficiently as possible.

- **Automated and Simplified Resource Integration:** Automation of the resource integration process, including the integration of fairsharing.io for metadata completion.
- **The APICURON Implementation Study:** A comprehensive implementation study to validate the enhanced platform's robustness and scalability. This involved the successful technical integration of APICURON with major European Bioinformatics Institute (EBI) resources.
- **Legacy Migration and Workflow Integration:** Engineered the migration of the DisProt database from a legacy system to a modernized architecture. This transition allowed DisProt to adopt a modern submission approach to APICURON and served as a key use-case to demonstrate how the enhanced platform facilitates credit attribution within complex curation workflows.
- **Privacy and Compliance Infrastructure:** In response to GDPR challenges identified during the implementation study, this thesis introduces the APICURON Consent Management Platform. Designed as a deployable infrastructure component for partner platforms, it streamlines GDPR compliance and securely manages user consent for data sharing.
- **Increased feature-set:** APICURON now provides additional features such as periodic leaderboards, privacy handling, and recognition group isolation, enabling adaptation to a variety of use cases when recognizing research assets.
- **Domain Expansion to Research Software:** Building on the framework's success in biocuration, this thesis extends APICURON's applicability to a new domain by conceptualizing a proof-of-concept for research software credit.

Part I

Credit & Recognition

Credit & Recognition

1.1 Shift Towards Data-centric Science

Recent high-profile breakthroughs, particularly in biology, have been directly attributed to data-driven approaches that leverage Artificial Intelligence and Large Language Models (LLMs)[17][40]. However, it is critical to recognise that these Artificial Intelligence (AI) systems are not isolated inventions or singular products. Rather, they represent the current apex of a cumulative technological stack, enabled by the democratisation of computing power through cloud infrastructure and the maturation of the Big Data phenomenon. AI is, in effect, the current “final node” in a long and complex chain of value production.

This production chain relies heavily on the availability of high-quality, structured data. In the scientific domain, this resource is produced, maintained, and refined by the intellectual labour of expert researchers. Yet, a critical misalignment exists within the academic ecosystem. While the scientific process has become data-reliant, the academic currency for excellence remains tethered to the static publication.

This discrepancy creates a system that rewards the final output, namely the literature or the AI model, while obscuring the foundational work of data curation that makes such outputs possible. This hidden labour is frequently undervalued, as traditional assessment practices often over-rely on a limited set of quantitative indicators, such as citation counts and the h-index. Despite the epistemological shift toward data-centric science, the mechanisms for acknowledging research outputs did not evolve in parallel. The current academic reward system suffers

from an observational bias.

To correct this disparity, current mechanisms of research assessment are currently undergoing a fundamental paradigm shift. Traditionally, assessment relied on proxy metrics derived from static publications (e.g., the h-index). However, the diversification of research activities has rendered these paper-based metrics insufficient. This misalignment has triggered a global reform movement, supported by key initiatives such as the Declaration on Research Assessment (DORA)[32] and the Coalition for Advancing Research Assessment (CoARA)[7]. These movements explicitly call for the acknowledgement of diverse outputs, including datasets, software, and curation work.

1.2 The Acknowledgement Pathway

In the landscape of research assessment, the notions of credit, recognition, and rewards are intricately connected, frequently appearing as terms whose definitions are either interchangeable or inconsistently applied across existing scholarly publications. Therefore, before examining the specific context of each, it is imperative to define these terms and explain how they are related in representing a pathway highlighting the process of acknowledging research contribution

Credit is defined as the formal assignment of roles and contributions to individuals involved in research activities or the creation of research outputs. This commonly takes the form of authorship in publications or formal acknowledgement for contributions to other outputs (e.g., research data, software) and activities (e.g., training). Standardised taxonomies, such as CRediT[25], have been developed to systematically describe each author's role in a research output.

Recognition represents the broader acknowledgement of a researcher's accomplishments and positive impact within their field. This extends beyond formal credit to include informal and community-wide forms of appreciation, such as awards, honours, citations, and public commendations. Recognition significantly enhances a researcher's reputation and career trajectory.

Rewards are typically tangible benefits received by researchers in exchange for their contributions. They serve to sustain motivation and provide the necessary resources for ongoing research. Examples include financial incentives, academic promotions, the securing of grants, and other forms of career advancement.

Therefore, the process of acknowledgement in research starts by formally **crediting** researchers and highlighting their involvement in the creation of research outputs. This credit progresses to the obtention of broader **recognition**, which in turn can be materialised in the form of **rewards**.

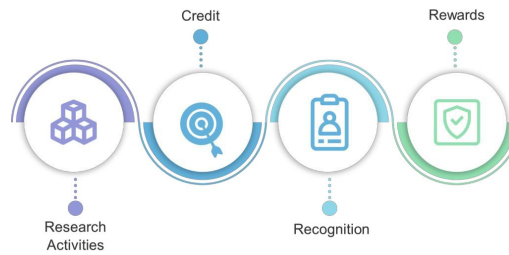


Figure 1.1: A diagram depicting how Credit, recognition, and awards form a pathway for acknowledging research activities.

1.3 Global Frameworks for Research Assessment Reform

The observational bias inherent in the current landscape has precipitated a crisis of confidence in research assessment practices. Consequently, the past decade has witnessed the emergence of high-level international initiatives aiming to reform how scientific excellence is evaluated. Collectively, these efforts articulate the community consensus and functional requirements that guide the technical solutions proposed in this thesis, explicitly calling for the recognition of diverse outputs such as research software and training activities.

1.3.a The Declaration On Research Assessment (DORA)

Established in 2012, the San Francisco Declaration on Research Assessment (DORA)[32] represents a foundational effort to improve the evaluation of scientific outputs. The initiative was driven by the need to assess research based on its intrinsic merit rather than relying on journal-based metrics, such as the Journal Impact Factor. DORA explicitly advocates for a broader range of assessment criteria, encouraging the utilization of diverse qualitative and quantitative measures to ensure fair evaluation.

Crucially for the scope of this work, DORA signatories support the adoption of inclusive assessment criteria that recognize and reward contributions to research software and training activities. In 2021, DORA expanded its scope by

launching Project TARA (Tools to Advance Research Assessment) to support the development of new policies. This effort included the introduction of Reformscape, a searchable collection of criteria for hiring, promotion, and tenure, designed to showcase global initiatives moving beyond traditional metrics.

1.3.b Principles for Responsible Metrics and Integrity

Following DORA, subsequent frameworks addressed specific aspects of the assessment problem, focusing on the responsible use of quantitative indicators. The Leiden Manifesto[16] established ten principles that serve as guidelines for quantitative evaluations in research assessment. It emphasizes that these evaluations should serve as support, not replacements, for informed judgment, advocating for the regular scrutiny and updating of indicators used, while considering the research context when assessing outputs. Furthermore, it safeguards differences in the research landscape, such as variations in publication and citation practices across different fields, and calls for the protection of excellence in locally relevant research

While the Leiden Manifesto refines the use of quantitative metrics, the Hong Kong Principles[23] focus on the underlying behaviors of research integrity. These principles offer a sharp critique of the current research assessment landscape, specifically targeting the exclusive reliance on publication counts, and Journal Impact factors. The principles highlight how the systemic incentives that prioritize novelty over rigor have resulted in distorted evidence.

The Hong Kong principles promote quality over quantity by directly rewarding responsible research practices such as meticulous data verification, effectively moving from impact metrics to integrity metrics. Crucially, the framework places a high value on complete reporting to mitigate reporting bias, ensuring that negative results are not obscured. Furthermore, the principles advocate for a research culture where Open Science[28] practices are encouraged and integral to the definition of excellence, alongside transparent assessment practices that acknowledge diverse facets of research output.

1.3.c The Coalition for Advancing Research Assessment (CoARA)

Perhaps the most significant recent development is the establishment of the Coalition for Advancing Research Assessment (CoARA) [7] in 2022, which brought together funding organizations, research institutions, and assessment authorities. The coalition is based on the Agreement on Reforming Research As-

essment, which establishes a common direction for research assessment reform.

The agreement outlines ten core commitments, the first of which is to “recognise the diversity of contributions to, and careers in, research”. Unlike previous general declarations, CoARA provides operational recognition mandates. It explicitly acknowledges that outputs beyond publications are valuable, citing the role of the software engineer as an indicative example of a role that must be recognized. Furthermore, CoARA emphasizes the importance of practices that contribute to robustness, openness, transparency, and inclusiveness. To facilitate implementation, the coalition has established working groups such as “Towards Open Infrastructures for Responsible Research Assessment” (OI4RRA), which focuses on the infrastructure required to support these reforms.

1.3.d Addressing the Invisible Labor: The Hidden REF

While DORA and CoARA drive reform from a top-down policy level, community-driven initiatives like The Hidden REF^[15] in the UK have emerged to challenge the invisibility of research roles from the bottom up.

The initiative was founded in response to the 2021 Research Excellence Framework (REF), where the vast majority of submissions were limited to journals and books, leaving other vital contributions unrecognized.

The Hidden REF is a volunteer-led campaign that organizes competitions specifically designed to highlight the large number of “invisible” contributions, including community engagement projects and open-source tools. Although not exclusively focused on software, the initiative actively supports the recognition of research software and training activities, frequently highlighting them as primary examples of the “hidden” work that underpins the research process.

1.3.e The Implementation Gap

Collectively, these initiatives demonstrate a global consensus: the research community must move beyond the use of simple publication metrics. However, despite the existence of these frameworks, significant limitations persist in their practical implementation. While there is a mandate to recognise diverse contributions, tracking specific types of research contributions, particularly software and training, faces technical limitations in existing platforms. There is a pressing need for enhanced standardisation in how these activities are cited and improved interoperability across credit platforms to operationalise these policy goals.

1.4 Standards and Taxonomies

While the policy landscape demands the recognition of diverse research outputs, the existence of robust metadata standards and taxonomies is paramount to reach a technical implementation consensus around credit & recognition. The community has developed specific frameworks to describe contributions to software, data, and workflows. However, a “granularity gap” remains between these static standards and the dynamic nature of these contributions. This section analyzes the current technical ecosystem, identifying where existing standards succeed and where they fail to capture the complexity of discrete research outputs and data-centric research.

1.4.a Contributor Roles and Taxonomies CRediT

The most widely adopted standard for attributing credit is the Contributor Roles Taxonomy (CRediT). CRediT aims to recognize the role of article contributors in a scientific contribution, shifting authorship from a simple list of names to a structured description of 14 specific roles, ranging from Conceptualization and Data Curation to Software and Methodology. Since its approval as an NISO ANSI standard, it has been adopted by over two dozen publishers to distinguish contributions in scientific publications.

While effective for traditional publications, CRediT is static as it describes the state of the research output and therefore operates at a coarse granularity, often insufficient for continuous, temporal research outputs; it assigns a static label to a contributor for an entire project but fails to quantify the magnitude or specific nature of the contribution. Where a researcher may contribute hundreds of distinct annotations over years, a single “Data Curation” label on a paper fails to capture the depth and effort of the work, particularly in fields like biocuration.

To address this lack of specificity, the Contributor Role Ontology (CRO)[8] was developed to extend CRediT. CRO provides finer-grained vocabulary designed to offer greater transparency and improved systems of attribution. Yet, even with the extended vocabularies, these taxonomies remain descriptive of outputs rather than processes, and do not solve the quantification problem, leaving a gap in tracking continuous research activities.

1.4.b Software Citation Standards

The research software community has developed standards to ensure software is treated as a primary research output. These efforts provide a relevant blueprint and baseline for crediting complex research outputs. Two key standards have emerged to support this.

- **Citation File Format (CFF)** A human- and machine-readable specification that developers add to code repositories to explicitly inform users how to cite the software. This promotes proper credit by lowering the barrier to accurate citation.
- **CodeMeta** A minimal metadata schema (JSON/XML) designed to standardize the exchange of software metadata across repositories and organizations. CodeMeta acts as a “concept vocabulary,” creating a crosswalk between different repositories to ensure interoperability.

These initiatives, supported by the FORCE11 Software Citation Principles[37], emphasize the importance of unique identification, persistence, and specificity in citation. However, standards like CFF and CodeMeta function primarily as discovery metadata, largely focused on the product rather than the effort, its valuation, and its assessment. While features like GitHub’s CODEOWNERS files in repositories can define responsibility for specific code sections, there remains a lack of standardized mechanisms to credit the process of development and maintenance over time.

1.4.c Provenance and granularity gap

The fundamental deficiency in current credit mechanisms is the focus on static deliverables rather than dynamic workflows. To accurately credit modern and complex research activities, a system must track the provenance of the output, the history of who completed which task, and when.

The W3C PROV Ontology[30] provides the foundational data model for representing granular contributions, defining entities, activities, and agents to represent individual atomic actions to an entity. However, raw provenance data is often too complex for academic recognition systems. While PROV-O provides the necessary precision for audit trails, it generates massive graph structures for trivial workflows. Leveraging this data for academic credit requires a semantic compression layer, an intermediary logic capable of abstracting low-level information into human-readable contribution metrics.

Conversely, simple metrics lack crucial precision. This creates what is known as the “Streetlight Effect”[12]: assessment frameworks prioritize what is easy to measure and track rather than what is actually important to the scientific process.

Current practices fail uniformly to recognize non-traditional outputs like curation and training, largely due to this lack of standardization in how usage and impact are tracked. While standards like SCoRO[35] (Scholarly Contributions and Roles Ontology) attempt to capture roles within specific timeframes, and CodeMeta standardizes software metadata, these solutions remain siloed.

1.4.d The need for an Aggregation Framework

The analysis of the state of the art reveals a granularity gap. The scientific community possesses high-level policies such as COARA and DORA demanding recognition for non-traditional research output, alongside low-level technical schemas and taxonomies such as PROV-O, CFF, and CRediT capable of describing said outputs. This is a data modelling trade-off. CRediT optimizes for human readability and summary, but lacks precision, while PROV-O optimizes for machine reproducibility and auditing precision, but lacks human readability.

This analysis reveals that the granularity gap is not merely a policy failure, but an architectural one. There is a lack of an intermediary logic layer capable of ingesting high-frequency provenance streams and dynamically aggregating them into high-value recognition tokens. Addressing this gap requires an inclusive framework that does not merely count static outputs but quantifies the continuous labour of research. This thesis proposes APICURON as that missing technical link, a platform designed to operationalize these standards into a cohesive credit and recognition ecosystem.

1.4.e Towards a Federated Ecosystem for Recognition

While the “granularity gap” highlights the need for a specialized processing engine for research activities, addressing the attribution challenge at a global scale cannot be achieved by an individual platform. The diversity of research outputs, that apply to various research communities ranging from software and datasets to training materials, demands a context-aware approach federated by different research communities. Therefore, a robust strategy for credit and recognition must be implemented not as a single tool, but as a federated ecosystem of interoperable components.

1.4.f The Architectural Components

This thesis proposes a conceptual architecture consisting of four distinct, interacting components to implement the high-level recommendations of CoARA and DORA. This comprehensive overview of the ecosystem guarantees a sustainable and context-aware approach to crediting from the source of contribution to the researcher's permanent record.

- **Research Publishing Platforms**

The foundation of the ecosystem consists of the platforms where research assets are created, stored, and maintained. These include research asset repositories (e.g., Zenodo[10], GitHub), biological databases (e.g., DisProt[27]), and activity registries (e.g., ELIXIR TeSS[4]).

In the context of recognition, these platforms act as the primary contribution event generators, capturing the initial metadata of a scientific contribution.

- **Researcher PID Providers**

Identity management is delegated to specialized providers that issue Persistent Identifier (PID). Services like Open Researcher and Contributor ID (ORCID) ensure the disambiguation of researchers across distributed systems. By incorporating these PIDs into metadata at the publishing layer, the ecosystem ensures that activities are linked to a verified identity rather than an ambiguous string name.

- **Research Metadata Aggregators & Knowledge Graphs**

To avoid fragmentation, metadata from disparate publishing platforms must be harmonized. Scientific Knowledge Graphs, serve as universal aggregators, collecting, cleaning, and deduplicating metadata to create a connected graph of research entities.

- **Researcher Contribution & Recognition Platforms**

This component focuses on the interpretation of contribution data. It transforms raw contribution metadata into meaningful recognition, considering the semantic depth of the contributions, providing adequate recognition to researchers. This includes Contribution Trackers which maintain detailed logs of researcher activities, and Academic Profile Platforms, which visualize these contributions to support career progression and recruitment.

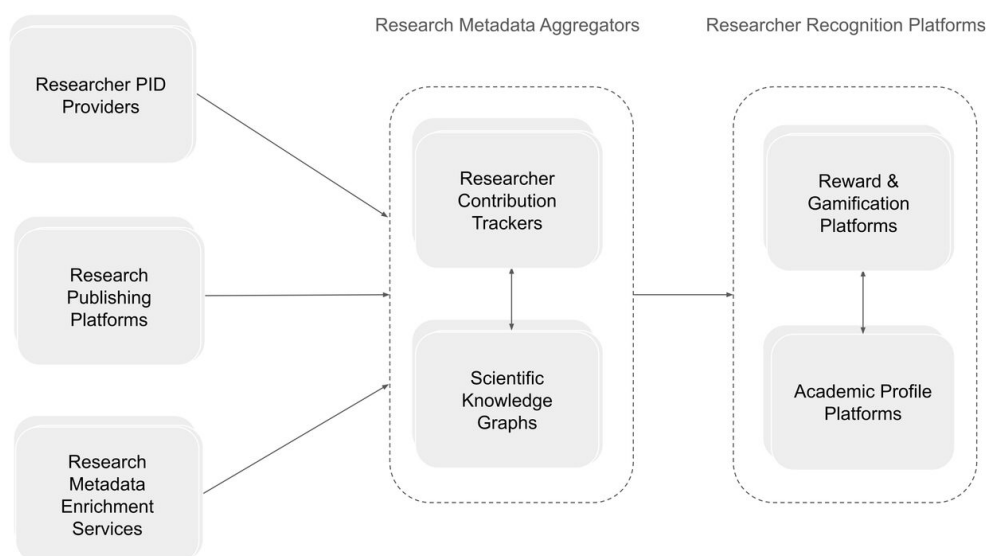


Figure 1.2: Bird's-eye view of the ecosystem for crediting shows the interactions among the various types of components. Adapted from author's work in ELIXIM3.1 Strategy for inclusion of credit for research assets decided

1.4.g Strategic Placement of APICURON

Within this federated topology, APICURON fills an architectural niche. It acts as a semantic middleware. While global Knowledge Graphs optimize for breadth and connectivity, APICURON acts as a semantic compression engine: ingesting raw provenance logs from Publishing Platforms and synthesizing them into verifiable, high-value, context-aware recognition assertions. This process allows the system to bridge the granularity gap by finally pushing these verifiable contribution assertions to PID Providers such as ORCID and Academic Profiles.

By situating APICURON within this hierarchy, this work moves beyond the concept of a standalone system. Instead, it proposes a modular recognition engine designed to bridge the gap between raw research logs and the formal academic record, enabling a flow of credit that is both granular and globally interoperable.

Part II

APICURON

APICURON is a platform specifically designed to credit and acknowledge scientific contributions beyond traditional academic indicators. It focuses on essential yet often overlooked outputs, such as biocuration, curated datasets, training materials, research software, and workflows. By integrating with research infrastructures, the platform tracks and aggregates these contributor activities, providing a standardized and verifiable mechanism for recognition through integration with researcher identifier providers such as ORCID. Consequently, APICURON promotes a broader, more inclusive view of scientific impact, empowering institutions and researchers by formally attributing the significant effort involved in collecting, curating, and processing scientific data.

In its mission to provide comprehensive credit and recognition, APICURON serves two primary stakeholders: research publishing platforms (e.g., research asset repositories and biological databases) responsible for curating and disseminating research, and the individual scientists who contribute their work to these platforms. For publishing platforms, APICURON offers the ability to evaluate contributors' engagement in an open science context. This is achieved via an integration process where platforms can start by:

1. Defining a Recognition Model on APICURON: This model specifies which research activities are recognized and configures associated reward mechanisms (e.g., badges, medals) based on the defined research activities.
2. Submitting contribution data mapped to the previously defined research activities, attaching a reference to the contributor, the type of contribution, and the time at which it occurred.

2.1 The Recognition Framework

APICURON operates on a flexible credit system designed to track and acknowledge contributions across diverse scientific domains, from biocuration to open-source software development. The foundation of this system is the Activity Definition. An activity represents a specific, credited action, such as “contributing to documentation” or “reviewing a publication” assigned a specific weight to reflect its relative importance or impact within a community.

Leveraging activity definitions, the system relies on Activity Reports. These reports serve as the granular records of activities, linking three critical entities: the agent researcher performing the action, the specific activity type, and the target object of the contribution, the research asset. This structure is based on the PROV-O[30, 38] entities and properties to represent a single activity occurrence.

To translate granular tracking into meaningful recognition, APICURON processes contribution reports to produce various outputs as illustrated in 2.1

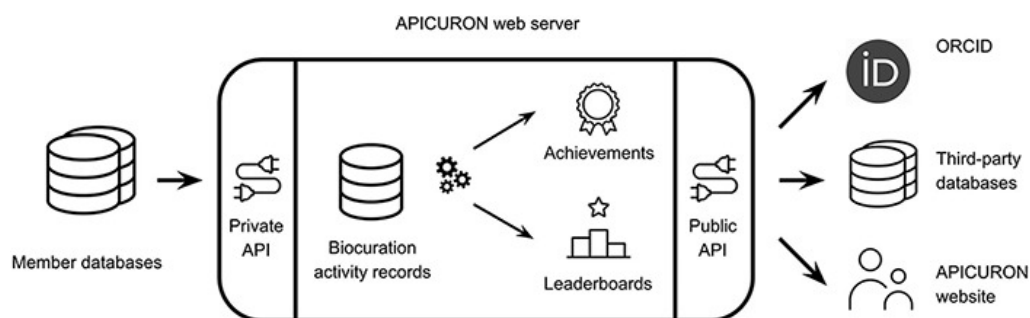


Figure 2.1: APICURON Workflow, highlighting the inputs, aggregation process, and products of the recognition system

2.1.a Metrics, Engagement, and Incentive Mechanisms

The system continuously aggregates incoming reports to generate contribution metrics. These outputs serve a dual purpose: providing platforms with feedback on scientific engagement, and offering users a comparative context for their contributions.

These metrics are strictly scoped to the intersection of a user and a partner resource. From a functional perspective, every measured indicator in the platform is identified by the unique combination of a researcher and a research platform.

This scoping is critical because Recognition Models are heterogeneous across resources; one platform may weight activities based on curation effort, while another prioritizes the scientific impact of the entity.

Consequently, cross-platform comparison is semantically invalid, and researchers must be evaluated solely within the context of the specific Recognition Model they engaged with. This design acknowledges that indicators across varying research infrastructures are incommensurable due to differing granularities, semantics, and weighting philosophies.

Temporal Segmentation

Furthermore, these metrics support flexible time windows. By adding a temporal dimension (e.g., weekly, monthly) to the user-resource pair, the system addresses the problem of ranking stagnation. This allows research publishing platforms to curate specific engagement strategies such as highlighting recent activity to mitigate inertia and foster a dynamic, inclusive environment

Comparative Context

To ensure contributions are recognized beyond the platform's boundaries, API-CURON facilitates the external dissemination of credit. The system synchronizes verifiable contribution data directly to researcher identifier providers (e.g., ORCID), which prevents credit from being siloed. This mechanism ensures that granular contributions become an integral part of the researcher's persistent academic record and professional portfolio.

Badges

Designed as user-centered achievements, badges appeal to the user's intrinsic motivation by offering a framework to visualize personal progress and enable consistency. Managed by the scientific resource, badges support continuous updates and progress tracking. They can also be configured with specific date bounds, introducing a time constraint that creates a sense of urgency, encouraging consistent engagement over time.

Medals

In contrast to badges, medals are designed to appeal to the extrinsic competitive drivers of the end-user. These are ranking-based achievements where users compete for specific standings (e.g., 1st, 2nd, or nth place) based on criteria defined by the scientific resource. Like badges, medals support time constraints

but are specifically configured to foster grouping. When well-configured, this grouping prevents user dejection by ensuring users compete within appropriate tiers rather than against an insurmountable global leaderboard.

Contextual Fairness via Contribution Groups

A critical challenge in the design of credit & recognition research infrastructure is the disparity between different classes of contributors. By quantifying labour and publishing performance indicators, and reward mechanisms, a system inherently constructs a hierarchy of performance that invites observation and comparison.

Comparing the high-volume output of professional, often-salaried curators directly against the sporadic contributions of community volunteers creates a metric misalignment. It presents a distorted view of reality where valid voluntary efforts appear negligible, potentially leading to the demotivation of community contributors or the misrepresentation of professional curators' impact.

To mitigate this, APICURON introduces Contribution Groups, a mechanism allowing research platforms to define logical cohorts to isolate metric aggregation. Each cohort can benefit from a unique recognition model and associated reward mechanisms. This ensures that competition and ranking occur only among peers with comparable resources and time commitments.

Moreover, the implementation of this feature avoids the pitfalls of static user tagging. Users are not assigned a group directly; Instead, APICURON utilizes event-level contextualization. Group membership is captured at the moment of the contribution, effectively tagging the activity report rather than the user profile.

While the conceptual goal is to tag every contribution with its appropriate context, mandating that research platforms explicitly tag every individual API payload with a group identifier would impose an undue integration burden. To resolve this trade-off between granular tracking and ease of use, APICURON implements a State-Based Defaulting mechanism.

Administrators manage cohort assignments via the platform's user interface, effectively setting a User-Group Binding state. This configuration acts as a forward-looking directive: it establishes the default destination for all future contributions generated by that specific user. This approach provides the following benefits:

- **Operational Simplicity:** The system adheres to a “Convention over Configuration” philosophy. If an incoming contribution report lacks a specific group tag, the ingestion engine automatically resolves the user’s current binding and assigns the contribution to their configured default group.
- **Explicit Override:** For complex workflows where a single user contributes in multiple capacities simultaneously, the API allows the default binding to be overridden on a per-report basis.
- **Historical Integrity via Event-Level Tagging:** modifying a user’s default group binding via the UI does not retroactively alter their data. It merely redirects the stream of subsequent events to the new group.

This distinction is vital for preserving the historical record. If a researcher transitions roles, (e.g a “Community Volunteer” is hired as “Core Staff”) the administrator simply updates their default binding. All subsequent activity is credited to the “Core Staff” cohort, while the previous body of work remains correctly immutably associated with the “Volunteer” group. This approach, leveraging event-level tagging, ensures that a researcher’s indicator history accurately reflects their career trajectory without the data distortion inherent in static, global user tagging.

This logic ensures that a single researcher can accumulate distinct sets of metrics and rewards corresponding to different phases of their career. By preserving this historical context, the system prevents retroactive distortion of data.

2.2 Technical Implementation

At the core of APICURON’s main workflow is the continuous processing of contribution reports following an evolving Recognition Model to generate statistics, engage reward systems and propagate information to the user’s Profile and ORCID. Architecturally, APICURON adopts a Three-Tier Architecture:

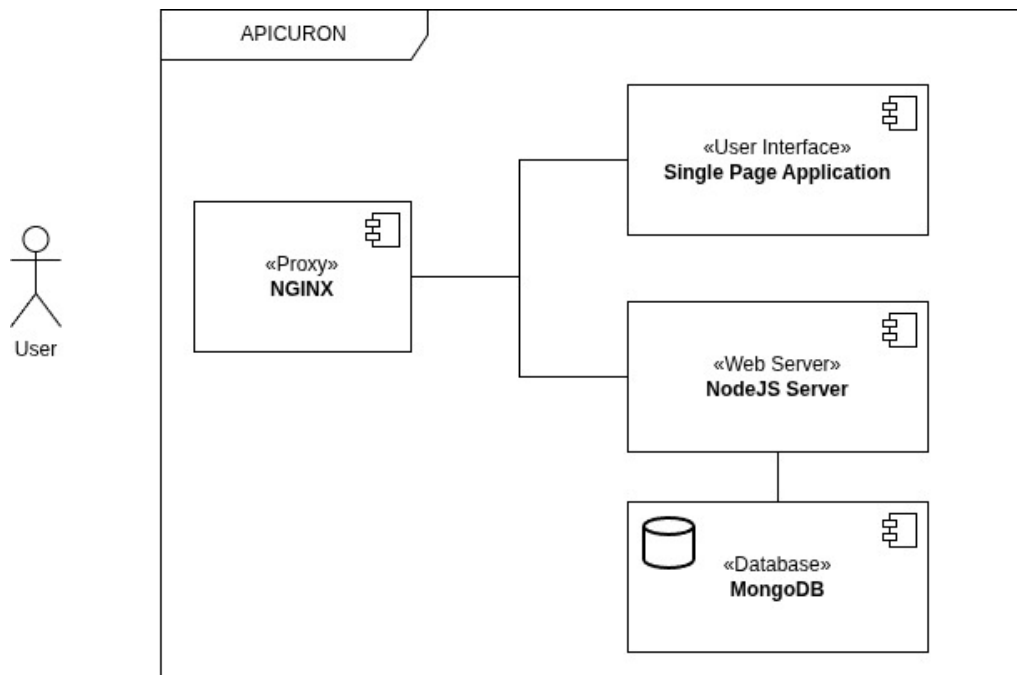


Figure 2.2: APICURON System architecture in UML Components Diagram

2.2.a Database Structure

APICURON utilizes MongoDB, a NoSQL database, to handle its persistence layer. Although the core entities (researchers, resources, and activities) exhibit relational characteristics, a document-oriented approach was prioritized to address two specific system constraints that a rigid relational schema would hinder.

First, the volatility of Recognition Models: As previously defined, Recognition Models determine how activities are weighted and rewarded. These models are subject to frequent iteration by the platform administrators. In a traditional SQL environment, reflecting these changes would necessitate frequent, computationally expensive schema migrations. MongoDB's flexible schema allows the system to adapt to evolving recognition definitions without incurring downtime or complex database refactoring.

Second, the optimization of read-time Latency: The platform's Presentation Layer relies heavily on real-time aggregated statistics, such as user leaderboards and profile summaries. Generating these views via normalization in a SQL database would require complex, multi-table JOIN operations that degrade performance at scale. By leveraging NoSQL's capacity for denormalization,

APICURON maintains pre-aggregated snapshots of data within the target documents. This ensures that high-volume read operations remain constant in complexity rather than scaling linearly with dataset size, effectively decoupling complex write-time processing from real-time user retrieval.

APICURON ensures a stable schema using an Object-Document Mapper (ODM) running in the application; this abstraction ensures the state of the database is compatible with the expected schema in the application layer. Given the use of an ODM, where database documents directly map to application objects, a Class Diagram effectively represents the data structure.

The schema employs a hybrid relational strategy. Relationships are reference-based (normalized) when strictly relying on identifiers, but utilize denormalization via document embedding when read-latency is critical. In these instances, a snapshot of the related object is embedded directly within the target document. Eventual data consistency is subsequently maintained through an event-driven propagation mechanism, ensuring that updates to the source entity are cascaded to all embedded instances depending on the consistency constraints.

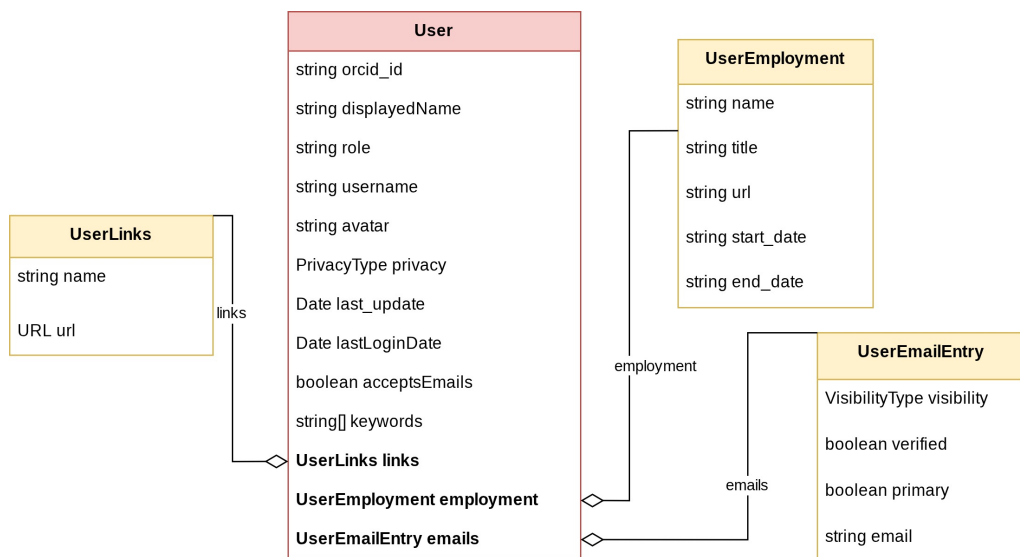


Figure 2.3: User schema design for APICURON. Classes in Red represent independent document collections. Classes in yellow are nested documents within the related Red classes

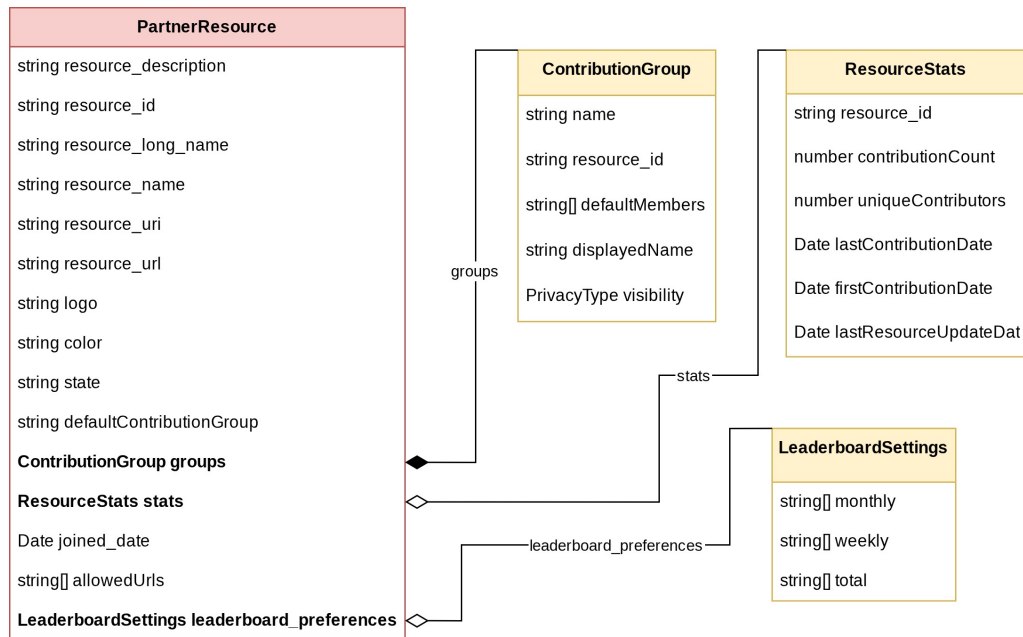


Figure 2.4: Resource schema design for APICURON. Classes in Red represent independent document collections. Classes in yellow are nested documents within the related Red classes

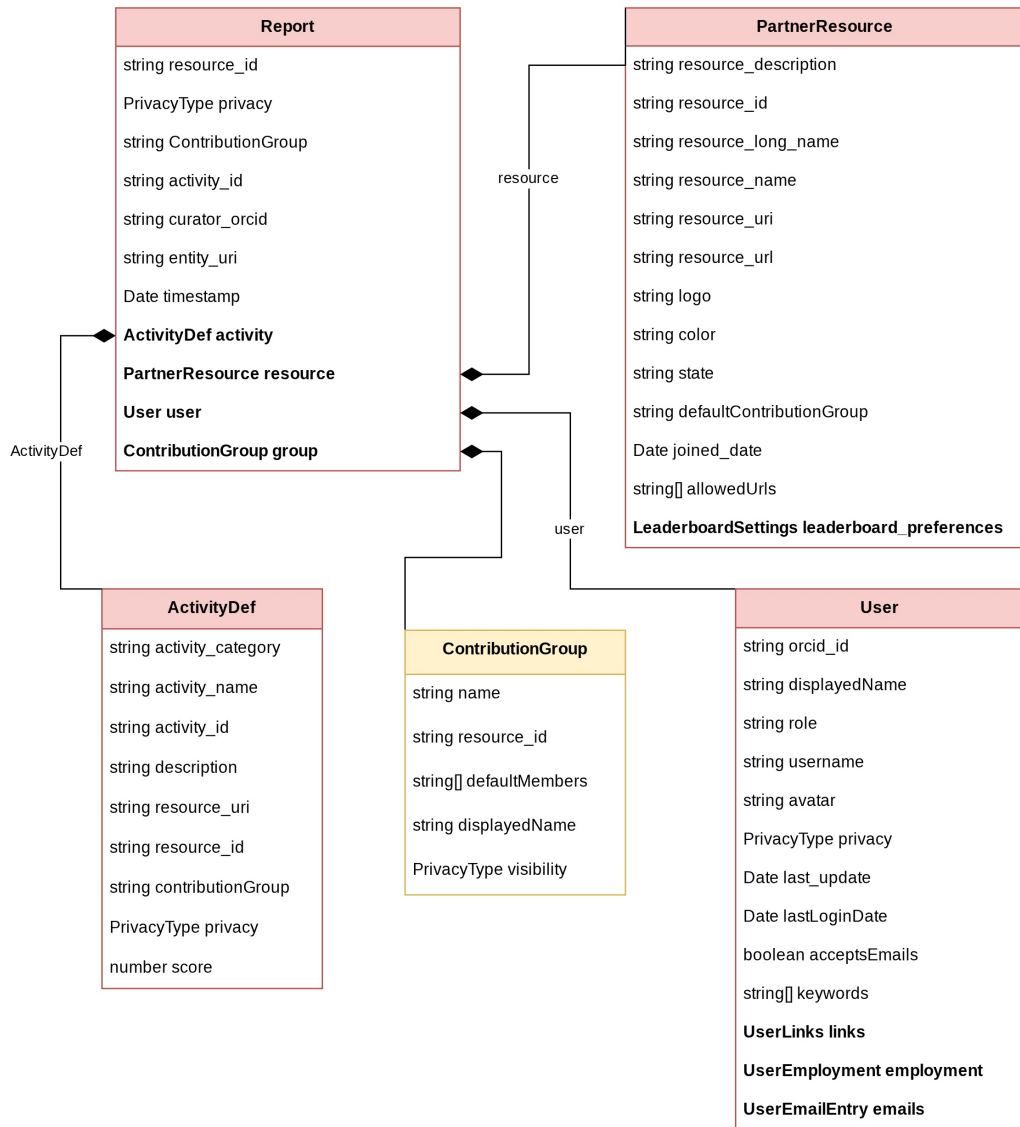


Figure 2.5: Contribution Report Schema design for APICURON. Classes in Red represent independent document collections. Classes in yellow are nested documents within the related Red classes.

2.2.b Authentication and Access Control

The APICURON API exposes public endpoints for general access, while sensitive routes require authentication and potentially additional role-based access control. This filtering is performed at the Hypertext Transfer Protocol (HTTP)[11] request level, where if not authorized, the request would be rejected and a response with HTTP code 401 Unauthorized would be issued.

Identity Management and Provenance

APICURON serves two primary stakeholders, comprising individual contributors and research platforms. To ensure data provenance and accountability, the system is designed such that every write operation, whether manual or automated, must be traceable to a human actor. Consequently, APICURON does not maintain separate “machine” identities for research platforms. Instead, authentication relies exclusively on the identity of the researchers managing those platforms.

Delegated Authentication via ORCID

The platform minimizes the risks associated with proprietary credential management by delegating identity verification to ORCID. Utilizing the OAuth 2.0 open authorization framework, users authenticate directly with ORCID, granting APICURON permission to verify their identity. This integration ensures that the system relies on a persistent, verifiable academic identifier rather than an isolated username-password pair.

While the web interface utilizes JSON Web Tokens (JWT) for session management, the data ingestion workflows often executed by automated scripts or legacy backend systems require a persistent authentication mechanism. To address this, APICURON implements a Personal Access Token i.e an Application Programming Interface (API) Key.

Authenticated users generate unique API keys via the user interface. Crucially, these keys are not generic “resource tokens” but are cryptographically linked to the specific user’s identity. When a script submits a payload using a key, the system resolves the key to the underlying user account. This mechanism ensures that even automated bulk submissions retain a clear chain of custody, identifying exactly which administrator authorized the data update.

Role-Based Access Control

With an authentication system exclusively tied to the user rather than the platform, access control is enforced through a strict Role-Based Access Control Role-

Based Access Control (RBAC) layer. When a request is received, the system first validates the user's identity and then queries the database to verify their specific relationship with the target resource (e.g., verifying that the user is a registered "Administrator" of the DisProt database). This decoupling of identity from permission allows for flexible governance, enabling resources to easily add or remove administrators without rotating global API keys.

2.2.c Data Ingestion

Unlike traditional aggregators that "pull" data, APICURON utilizes a push-based model. Research infrastructures integrate with the platform by sending contribution payloads via API calls, implementing a push-based data synchronization model.

APICURON provides a versatile set of REST API endpoints to accommodate the varying workflows of each research platform, from granular timely submissions via webhooks to asynchronous bulk transfers. Conversely, to accommodate platforms relying on scheduled submissions at fixed intervals, APICURON supports bulk ingestion workflows

The Timely Submission

APICURON offers two HTTP POST endpoints for a timely submission of activity reports on an "as-they-occur" basis. The `/reports` endpoint accepts an array of contribution events. This endpoint allows platforms to micro-batch data submissions into a small array that does not exceed 50 reports. This payload cap effectively acts as a micro-batching mechanism; payloads exceeding this size incur serialization and processing overhead that align better with the bulk submission approach. To optimize for the high-frequency nature of individual curation events, a dedicated `/reports/single` endpoint is exposed to bypass the overhead of array parsing and validation required by the micro-batch endpoint, additionally creating an endpoint that tolerates a higher concurrency rate when necessary.

The Bulk Submission

APICURON offers an approach for research platforms to submit or resubmit their entire log of activity reports periodically. This approach helps some platforms lacking an event-driven approach to compile the log of events periodically. The `/reports/bulk` HTTP POST endpoint accepts a JSON file containing a list of activity reports; this file can be large.

This endpoint offers a redundancy reduction strategy feature when parsing the

JSON file. Leveraging the default override pattern, “Global fields” whereby repeated fields in each activity report within the array can be specified only once as a default and can later be overridden on a case-by-case basis if necessary.

```
1 {
2   // Meta Header
3   "resource_id": "disprot", // Default Value
4   "reports": [
5     {
6       "activity_term": "region_first_generation",
7       "timestamp": "2016-08-11T09:59:17.000Z",
8       "curator_orcid": "0000-0002-0341-4888",
9       "entity_uri": "https://identifiers.org/disprot:DP00034"
10    },
11
12    {
13      "activity_term": "region_first_generation",
14      "timestamp": "2016-08-11T09:59:17.000Z",
15      "curator_orcid": "0000-0002-0341-4888",
16      "entity_uri": "https://identifiers.org/disprot:DP00034",
17      "resource_id": "another_resource" // override of the Default
18    }
19  ]
20 }
21
22
```

Evidently, the processing of this file is deferred, and the propagation of the results is not to be expected immediately, depending on the size of the data. This method implementation anticipates large amounts of data and applies optimizations in that regard.

Crucially, to ensure data durability within this asynchronous architecture, APICURON employs a Persistence-First ingestion strategy. Upon receipt, all valid contribution payloads are immediately committed to the Persistence Layer. Only after successful storage is the HTTP acknowledgment returned to the data provider and the event pushed to the in-memory processing queue. This guarantees that the volatile nature of the Node.js in-memory queue does not compromise data integrity; if the service fails immediately after ingestion, the raw record remains safely persisted.

In conclusion, by decoupling the ingestion interface from the processing logic, APICURON ensures that high-volume data submission does not block system resources. Once a payload is validated and ingested via this interface, it is

Method	Route	Description
POST	/reports	Micro-Batch Ingestion: Accepts up to 50 activity reports. Designed for platforms buffering events to reduce HTTP overhead without sacrificing near-real-time processing.
POST	reports/bulk	Asynchronous Bulk Ingestion: Accepts a complete dataset dump. Includes "Global Field" optimization to reduce payload size. Processing is deferred.
POST	reports/single	Real-time Ingestion: Processes a single atomic curation event. Optimized for low-latency, immediate consistency.

Table 2.1: Table Caption

immediately handed off to the asynchronous processing tier.

2.2.d The Event-Driven Processing Engine

While the Ingestion Interface handles entry, validation and dissemination features, these functions rely heavily on the core logic of APICURON, which resides in the Event-Driven Recognition Engine. This subsystem is responsible for the continuous, non-blocking execution of Recognition Models, metric aggregation, and reward propagation.

Currently, this processing pipeline is orchestrated via an internal task queue implemented directly within the Node.js runtime. This design choice leverages the platform's asynchronous, non-blocking nature to handle current event volumes with minimal infrastructure overhead while maintaining strict architectural decoupling. By isolating the event producers from the consumers, this in-memory queue serves as a transparent abstraction layer. This allows for a seamless future migration to a distributed message broker to support horizontal scaling without necessitating a rewrite of the core processing logic.

Unified Event Abstraction

Fundamentally, the outputs generated by APICURON are a function of two volatile, complex, and interdependent states: the continuous stream of incoming contribution events and the evolving definitions of the Recognition Models. A change in either the set of contribution events or the Recognition Model for a single research platform introduces a cascade of updates to synchronize the outputs APICURON produces. Additionally, the nature of these changes is incon-

sistent; changes range from continuous, singular updates to sporadic, massive dataset releases where data providers (research publishing platforms) refresh their full-set of contribution events. Conventional data-intensive architectures synchronize such dynamic variables through scheduled batch processing, where the entire dataset is periodically re-evaluated to ensure consistency. However, relying exclusively on this temporal approach introduces a significant inefficiency: it forces the system to blindly recalculate stable data to capture even a minority of changes, resulting in wasted computational cycles.

Despite this inefficiency, the system cannot simply discard bulk processing capabilities, as it must support those practical instances where data providers necessitate a complete dataset refresh. To reconcile this contradiction, APICURON fundamentally redefines how a state change is perceived. Instead of segregating inputs into streams for individual items and batches for bulk updates, the system formalizes every modification as a discrete event. Whether the input is a singular new contribution, a modification to a recognition rule, or a request to reload an entire repository, each is encapsulated as a standard event packet and placed into a unified processing queue.

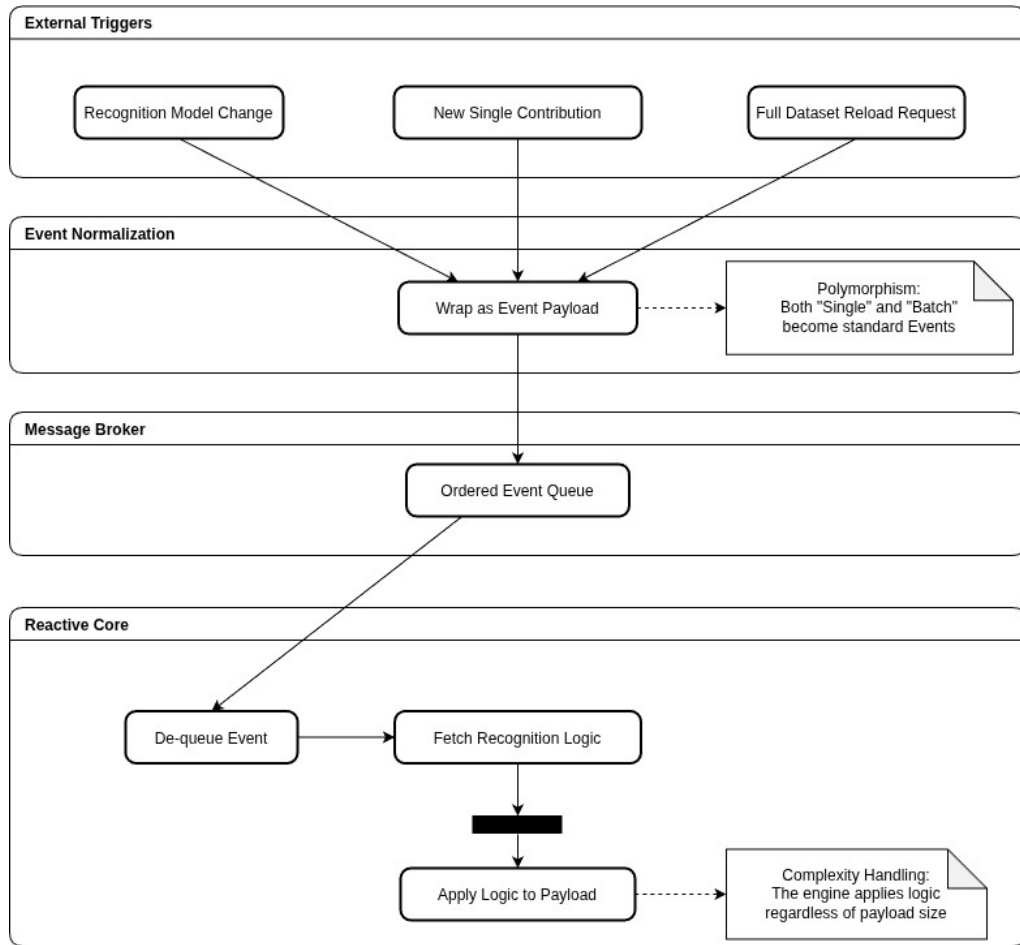


Figure 2.6: UML Activity Diagram showing the unified event pipeline demonstrating how disparate inputs (Single Contributions, Batch Reloads, Model Updates) are normalized into a standard event stream.

This operational shift establishes the foundation of APICURON’s event-driven architecture. By normalizing these disparate inputs, the system effectively abstracts the underlying processing paradigms; a massive batch is no longer a separate operational mode but simply a coarse-grained event with a larger payload. This unification ensures that the system remains purely reactive, triggered strictly by the arrival of an event rather than an arbitrary schedule.

This unification of *Stream* and *Batch* processing also provides a robust mechanism for Crash Recovery and State Reconciliation. Since the system is already designed to process large, coarse-grained Bulk events adapting to legacy constraints, this logic is repurposed to handle system restarts. In the event of a

service interruption where the volatile in-memory queue is lost, the system can leverage the persisted raw reports to establish a consistent baseline. On startup, the system identifies unprocessed records and re-ingests them using the standard Bulk event pipeline. Consequently, the legacy requirement for batch processing effectively functions as a native recovery strategy, eliminating the need for complex external transaction managers.

Consequently, the processing logic adapts dynamically to the specific type of event handled. As illustrated below, specific handlers are invoked whether the event is a new contribution, a badge definition update, or a medal ranking recalculation, ensuring that complexity is handled at the logic level rather than the infrastructure level.

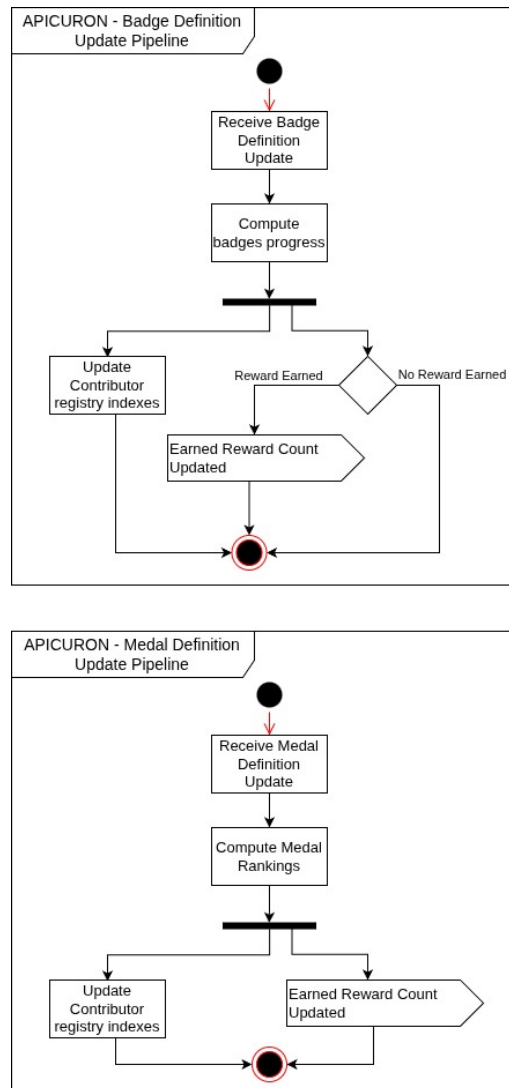


Figure 2.7: UML Activity Diagram showing the reward update pipeline for Medals and badges in APICURON

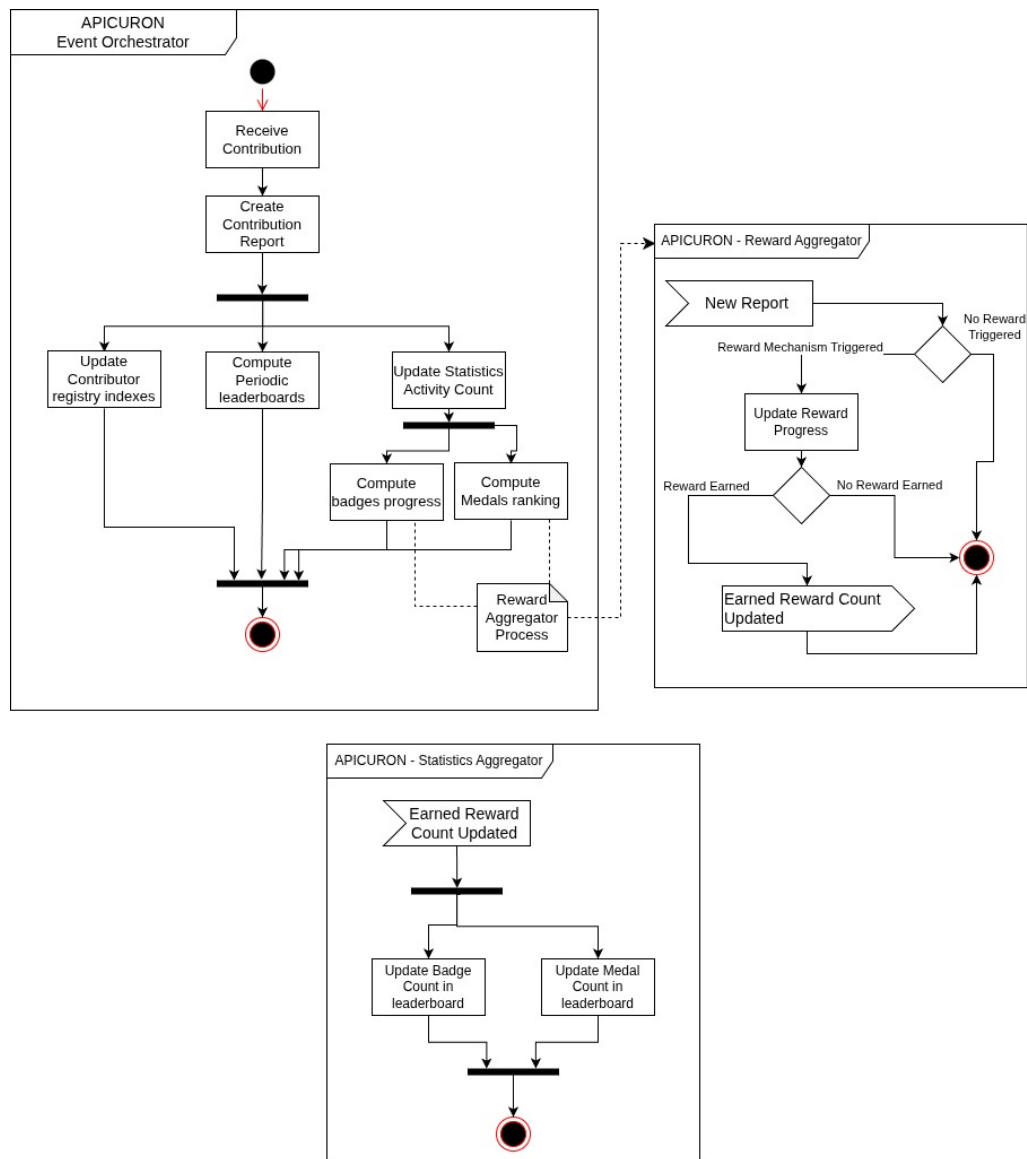


Figure 2.8: UML Activity Diagram showing the general aggregation pipeline and how it's linked to independent workflows enabling modularity

2.2.e Data Distribution & Dissemination

While the Ingestion Layer focuses on high-throughput data entry, the Distribution Layer is responsible for the contextualization and dissemination of these results. This layer operates on two distinct fronts: internal distribution, which segments data for accurate comparison within the APICURON interface, and external dissemination, which synchronizes credit with third-party researcher identities such as ORCID.

Internal distribution segments data for accurate comparison within the APICURON interface. This is performed by exposing endpoints leveraged by the Presentation Layer, the Single Page Application (SPA), to retrieve user profiles, leaderboards, and real-time statistics.

Beyond internal metrics, APICURON functions as a bridge to the wider academic ecosystem by propagating verifiable credits to third-party researcher identities, specifically ORCID. However, this integration presents two distinct challenges: an Impedance Mismatch regarding system reactivity and a Granularity Mismatch regarding data consumption.

The granularity mismatch

Raw biocuration events are often too granular to be meaningful on a professional résumé. A researcher contributing 500 individual annotation updates does not require 500 separate entries on their ORCID profile.

To resolve this, APICURON implements a semantic summarization layer. Instead of propagating raw event logs, the system synthesizes contributions into a coherent, human-readable narrative. Crucially, this synthesis is configurable by the research platforms and offers the following benefits:

- **Contextual Templating:** Research platforms can define custom text templates on a per-group basis. This allows for high-level distinctions; for example, a Senior Curator might receive a detailed citation format, while Community Volunteers receive a consolidated activity summary.
- **Payload Efficiency:** By collapsing thousands of micro-events into a single, updated summary object, the system significantly reduces payload size while maximizing the CV-readiness and professional value of the attribution.

Part III

DisProt

3.1 Operational Context & Legacy constraints

While APICURON provides the theoretical and architectural framework for granular credit, its efficacy can only be validated within a live, production-level research infrastructure. The Database of Protein Disorder was selected as the primary pilot implementation for this system. DisProt[27] is an open database integrating experimental evidence on intrinsically disordered proteins Intrinsically Disordered Proteinss (IDPs), intrinsically dis-ordered regions Intrinsically Disordered Regionss (IDRs), and their functions. Over the past two years, the database has grown over 20%, now comprising 3201 IDPs and 13347 pieces of evidence, including over 1500 new structural state annotations and more than 1300 new function annotations. Unlike automated repositories, DisProt relies entirely on the intellectual labour of domain experts to extract complex experimental evidence from literature. This curation process exemplifies the invisible labour phenomenon often observed in data-centric biology. While the resulting structured data is highly cited, the granular efforts of extraction defining region boundaries, validating evidence, and resolving conflicts remain opaque in traditional citation models.

Prior to this implementation, DisProt functioned as a traditional static database. Contributors were acknowledged only via generic “consortium authorship” on periodic update papers. This model, although effective, fails to capture the velocity and specificity of individual contributions. The integration of API-CURON aimed to transform DisProt from a static repository into a dynamic, credit-generating system.

3.1.a Semantic Integration Challenges

The primary challenge in integrating DisProt with APICURON was the semantic mapping of complex biological workflows into the standardized Activity Report schema described in [chapter 2](#). DisProt curation is not a singular action but a composite workflow involving:

- **Entity Identification:** Defining region boundaries within a specific identifiable protein (e.g., residues 1–213 in protein with ID: DP00287).
- **Ontological Annotation:** Assigning classes from the Intrinsically Disordered Proteins Ontology (IDPO)[[27](#)] or GO[[2](#)] ontologies describing the structural states, transitions, and protein functions to the target regions.
- **Evidence Validation:** Linking assertions to experimental methods (e.g., NMR, CD) and literature citations.

Given the complexity of the contribution workflow, integrating APICURON was not merely a matter of connecting API endpoints; it required a semantic translation layer.

3.1.b Decoupling Execution from Attribution

Legacy DisProt was designed as a traditional Create, Read, Update, Delete (CRUD) application. In this model, the system’s primary objective is State Persistence, fitting the primary purpose of the platform. The system focuses on overwriting the record upon requested changes. However, APICURON operates on a contribution model, focusing on the contribution action itself rather than the final product. Fundamentally, the legacy system prioritized state persistence, answering the query “What is the final state of this object?” In contrast, the credit attribution model requires a delta analysis, answering “What is the semantic distance between State A and State B, and what human effort was required to traverse it?”.

A simple “Update” operation in a CRUD system is opaque; it collapses high-effort intellectual contributions (e.g., defining a new annotation) and low-effort maintenance (e.g., fixing a typo) into a single technical transaction.

3.1.c Legacy Modernization Strategy

Considering the importance of DisProt within the IDP community, the technical debt incurred within the backend setup and the required change in processing approach. Given the technical debt and the imperative for high availability, a full rewrite was deemed unsustainable. Instead, an iterative migration strategy

was adopted.

To resolve the technical debt without a full-rewrite, the strangler-fig pattern[24] is used to introduce a new backend component in the system. This allows it to move away from the legacy Restify[34] web-server system to a newer, more flexible environment that is still Node.js[26] based but offers additional flexibility. The new component provides a baseline to introduce additional processing logic, and the APICURON integration using the strangler-fig pattern[24].

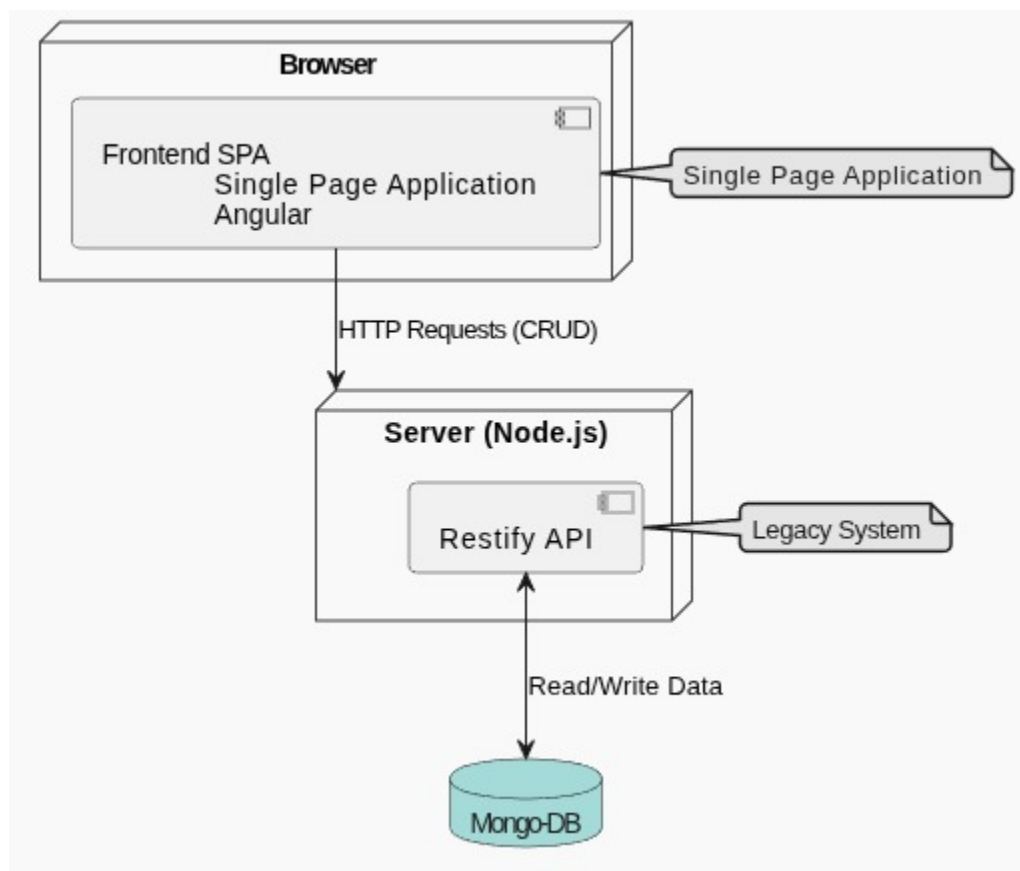


Figure 3.1: UML Component diagram showing the legacy DisProt architecture with a singular monolithic backend connected solely to the document database

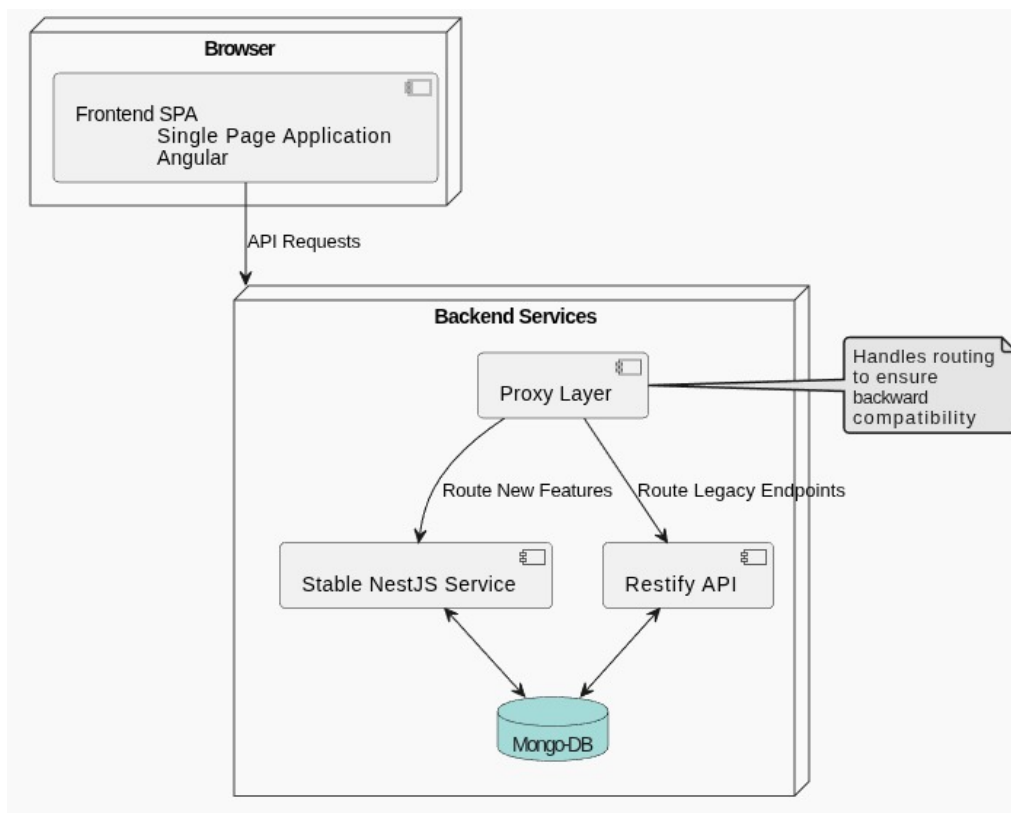


Figure 3.2: UML Component diagram showing the improved DisProt architecture implementing the strangler-fig pattern with two backend servers connected solely to the document database and a proxy to manage request routing

Although a new environment for implementing logic is provided, it is trivial to understand that the backend system only captures CRUD events, the endpoints that the DisProt frontend uses remain CRUD based. Therefore, a system to convert the opaque CRUD operations to contribution events is necessary. Crucially, the maintainability of this system is of the highest importance, as change is frequent and is paramount for the credit mechanism. To overcome the rigidity of the legacy CRUD architecture, the system was migrated to an Event-Driven pattern. A core objective of this migration is to decouple the execution of curation tasks from the interpretation of their value.

In a traditional approach, the interpretation of complex biological data entry through CRUD requires intertwined evaluation logic. Handling this imperatively would require a nested conditional structure. e.g: creating a new Disordered Region in DisProt might simultaneously imply:

- The introduction of a novel paper to the database.

- The completion of high-quality Minimum Information About Disorder Experiments (MIADE)[20] fields.

An imperative implementation of this logic would be a series of nested if-statements, which violates the Open-Closed Principle[21]; adding a new credit criterion in the future would require modifying the core persistence function, risking regression errors.

To overcome the limitations of imperative nesting, the system implements a Hierarchical Event Propagation model. This architecture inverts the control flow: instead of the persistence layer explicitly invoking credit logic, it broadcasts immutable Semantic State Changes. Formally, this implementation leverages the hybrid of the observer pattern and composite pattern. We define the contribution process as a directed acyclic graph where:

Here is the conversion of your logic flow into valid LaTeX for the mathematical components. I've kept the structure clean so the relationships between the entities remain clear.

- The Subject ($E_0 \dots E_n$): The root nodes represent the atomic user actions (e.g., `region_defined`).
- The Payload Constructors ($PC_1 \dots PC_n$): The first layer of logic handlers is to load the payload and compute the change delta for further processing. These handlers represent a function: $PC_n(p) = S$ where p is the atomic user action payload, and S is the computed delta and payload highlighting the meaningful change in state.
- The Payload Analysis ($PA_1 \dots PA_n$): Event handlers that are triggered by the payload constructors; they analyze the change in data to identify creditable activity. Each change of interest is checked and validated in this process before triggering the creditors.
- Creditors ($C_1 \dots C_n$): Creditors directly call the APICURON API. They represent the final layer in identifying activity.

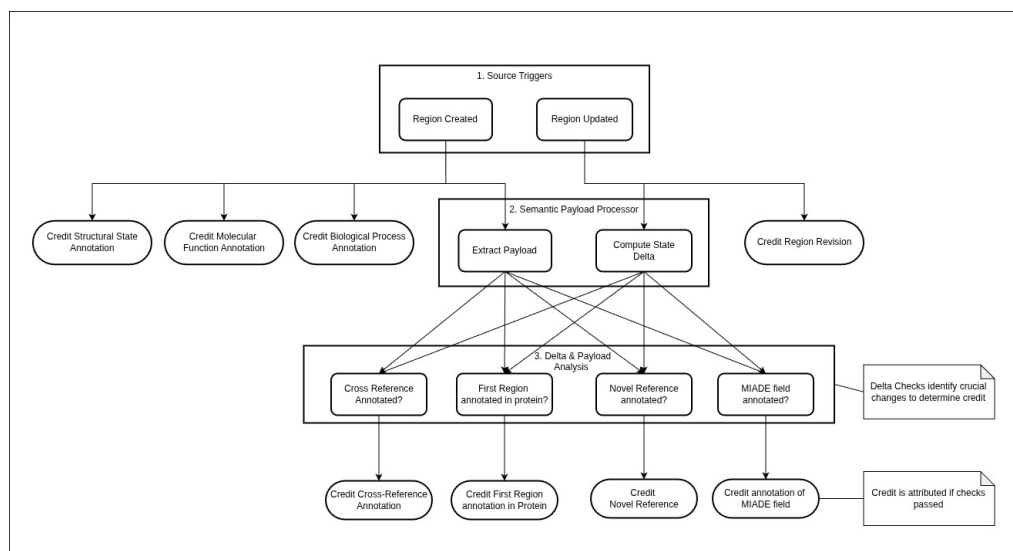


Figure 3.3: Event tree diagram depicting the semantic event propagation from atomic user actions to crediting operations

This ensures strict adherence to the Open-Closed Principle: new attribution logic can be grafted onto the event tree without modifying the underlying legacy persistence logic. Effectively offering optimal logic reuse.

3.2 The Deposition Process

To increase community engagement and gather additional scientific evidence for curation, DisProt introduced the article deposition page. A simplified form designed for publication authors and non-expert curators which allows users to contribute their experimental evidence and reference whether published or under review. This feature effectively reduces the barrier to entry for researchers who are unfamiliar with the complexities of scientific data curation practices, turning DisProt into a hybrid resource that serves both as a knowledge-base and as a deposition platform. Deposited data is reviewed by expert curators before publishing findings to maintain high-quality standards.

To facilitate this expansion, DisProt required a mechanism to capture contributions from the broader community, not just internal staff. This necessitated the development of the Article Deposition Page, a dedicated frontend interface designed to be "APICURON-aware." Technically, this interface acts as a provenance sensor. When an external user submits data via the deposition form:

- The user authenticates via ORCID

- The interface captures the submission metadata (PubMed ID, UniProt ID, experimental details).
- Upon validation by an expert curator, the DisProt backend triggers a POST /reports event to the APICURON API.

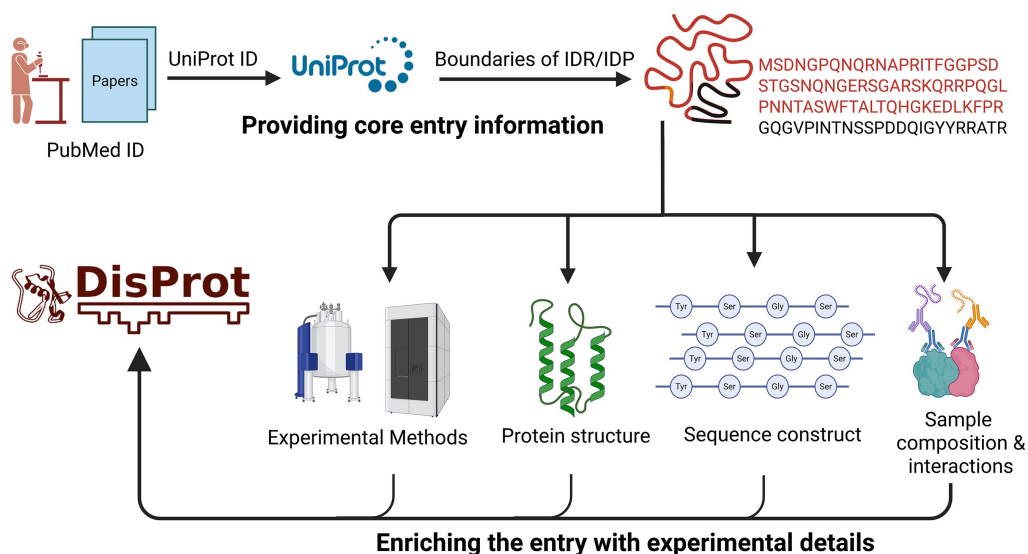


Figure 3.4: Workflow describing the important steps in the article deposition process. Created in BioRender. Aspromonte, M. (2025) <https://BioRender.com/z5rq1gc>.

The Deposition Page has transformed the resource into a hybrid knowledge-base and deposition system, integrating it with ELIXIR’s Deposition Database collection. These selected databases meet the technical quality and governance criteria for ELIXIR Core Data Resources and align with FAIR principles.

3.3 Operational Recognition

Upon the activation of the event stream, DisProt utilized APICURON’s feedback loops to visualize and quantify labour that was historically opaque. While the “consortium authorship” model remains in effect for high-level acknowledgment, APICURON was deployed to quantify curation efforts granularly.

The crediting instruments were calibrated to steer community behaviour toward specific quality standards, functioning as a “nudge” mechanism. This included:

- **Weighting Evidence Precision:** The scoring logic applied a higher weighting to structural evidence derived from high-precision experimental meth-

ods, thereby devaluing assertions based on lower-resolution techniques (e.g., X-ray crystallography missing residues).

- **Incentivizing Metadata Completeness:** Additional credit was algorithmically assigned for the annotation of MIADE compliant fields, encouraging adherence to community standards.

Furthermore, crediting mechanisms were integrated to target intrinsic motivational drivers. Temporal badges were implemented to visualize progress toward defined competency milestones. Concurrently, dynamic leaderboards were utilized to prevent ranking stagnation; by employing temporal segmentation, the system highlighted recent activity rather than cumulative historical dominance, thereby maintaining competitiveness for new contributors.

3.4 Impact Analysis

The integration of APICURON into the DisProt ecosystem provided the first opportunity to rigorously quantify the “invisible labour” of biocuration using longitudinal data. This analysis leverages time-series data comprising over 52,000 curation reports collected from the DisProt database.

The implementation of the Deposition Page successfully lowered the barrier to entry for external contributors. In the 2026 release, the database expanded by 20%, now housing 3,201 IDPs and over 13,000 evidence items. Notably, the provenance tracking capabilities of APICURON revealed that 19% of new entries and 46.5% of new annotations were generated by external contributors. This validates the efficacy of the “Deposition Page as a mechanism for capturing community input, effectively managing the impedance mismatch between casual contribution and rigorous curation standards.

Part IV

The APICURON Implementation Study

The APICURON Implementation Study

Having defined the architectural specifications and data schema of APICURON in the previous chapter, it is necessary to validate these models within the heterogeneity of the real-world biocuration ecosystem. A theoretical framework for credit attribution is only as valuable as its ability to integrate with existing infrastructure and align with the diverse workflows of working scientists.

To achieve this, an implementation study was conducted during the initial deployment phase. A dedicated cohort of early adopters was engaged to stress-test the system's interoperability and social acceptance. This group was specifically chosen to represent a diverse array of use cases, reflecting the broad spectrum of activities involved in biocuration ranging from the granular, manual annotation of intrinsic disorder (DisProt[27]) to the high-throughput classification of protein families (Pfam[13]).

The feedback and practical friction points encountered during this study were instrumental in transitioning APICURON from a prototype to a robust service. This chapter details that validation process. It begins by outlining the engagement methodology, followed by an analysis of the specific integration strategies required for each partner resource. It concludes by evaluating the impact of these integrations and, crucially, analyzing the unforeseen challenges regarding privacy and curator anonymity that necessitated the architectural evolutions discussed in later chapters.

4.1 Objectives and scope

The primary objective of this implementation study was to validate the APICURON framework across a heterogeneous landscape of biological databases. Rather than a monolithic deployment, the scope of this study was designed to assess the system’s adaptability across three critical dimensions

- **Workflow Granularity:** Testing the schema’s ability to capture both high-effort, low-volume manual curation (e.g., DisProt[27], Reactome[22]) and low-effort, high-volume curation such as Pfam[13].
- **Credit Attribution Models:** Evaluating the efficacy of different recognition mechanics, from granular activity tracking to competitive incentives and peer-review acknowledgment.
- **Technical Interoperability:** assessing the robustness of the API integration across distinct software architectures and development stacks.

This foundational study aimed to transition APICURON from a theoretical model to a production-grade infrastructure, identifying friction points in both technical scalability and social acceptance among professional curators.

4.2 Methodology

The integration of APICURON into the biocuration ecosystem was not merely a technical deployment but a sociotechnical undertaking. It required establishing community consensus on the definition of credit before any software could be effectively utilized. The implementation methodology was therefore bipartite with each scientific resource, consisting of a community engagement phase to define semantic standards and a technical execution phase to ensure seamless interoperability.

Following the consensus building and integration of APICURON. An implementation effectiveness analysis is performed. The purpose of the analysis was to determine the overall effectiveness of APICURON. The overarching goal is the understanding of the effectiveness of the acknowledgement provided through the implementation. The project took into consideration APICURON and all the interconnected databases, using a two-pronged approach: a qualitative, and quantitative analysis.

4.2.a Implementation

To avoid the pitfalls of a top-down schema design, which often leads to poor adoption in heterogeneous systems, a participatory design approach was employed. This phase focused on establishing a controlled vocabulary for “credit-worthy” activities that could satisfy the distinct granularity requirements of the partner resources. The consensus was derived through a structured analysis of four critical dimensions for each member of the early adopter cohort:

- **Strategic Alignment:**
This dimension evaluated the specific strategic goals of each resource (e.g., improving data quality). By mapping these objectives to existing deficits in the current acknowledgement systems, specific research activities can be prioritized. This configuration allows resources to leverage APICURON to incentivize community effort on critical tasks by ensuring they receive immediate visibility and recognition.
- **Attribution Granularity:**
This assessment addressed the disparity between the intellectual effort required for a task and the visibility of the reward. The objective is to recalibrate the recognition model to ensure that high effort contributions are adequately recognized in comparison to the low effort contributions counterparts.
- **Contextual Fairness:**
This dimension addressed the challenge of metric distortion caused by flattening distinct functional roles into a single leaderboard. The analysis distinguished between disparate user groups and their role within the scientific resource. e.g the effort of a standard contributor differs fundamentally from the editorial responsibility of an expert reviewer. This differentiation was essential to ensure a “fair comparison” environment, preventing the inequitable juxtaposition of varying user profiles creating a false equivalence.
- **Technical compatibility:**
This technical assessment evaluated the legacy architecture and update frequency of the partner resource. It determined the feasibility of real-time synchronization versus bulk-data ingestion, defining the integration strategy for each partner.

4.2.b Concepts

Before the results of the implementation study can be analyzed, it is necessary to present the relevant concepts and theory. The analysis considers the following

precepts and factors:

- **Knowing the User:** who is the user and how do they like to engage?
- **Is it worth it:** Professional vs non-professional curator value (perceived usefulness) considerations
- **Motivation:** Intrinsic vs extrinsic motivators
- **Rewards:** Are they a sufficient and effective motivation?
- **Effectiveness:** Does crediting improve curation quantity or curation quality or both?
- **Measurement:** How to measure and ensure crediting impact?
- **Consistency:** Do the results of the different measurement (quantitative or qualitative analysis) techniques agree?

In order to understand the user, some Psychology and Engagement concepts were considered. The first was motivation which refers to the psychological processes for initiating and continuing goal directed behaviours. Crediting is designed to foster motivation so it is important to align elements with user motivation. Additionally the Self Deterministic Theory (SDT) was also utilised. It is based on the idea that people are inherently motivated certain basic psychological needs eg. Autonomy, Competence and Relatedness as a function of intrinsic motivation which involves engaging in activities for their enjoyment rather than for external rewards or pressures (extrinsic motivators). The ideal scenario in crediting would be to align design elements with intrinsic motivators in order to improve engagement by leveraging inherent user qualities.

These considerations and theoretical underpinnings highlight the importance of the design aspect and how it must be informed by the user type in order to ensure a maximal experience by catering to the specific needs shown by the user. General game theory suggests[39] there are six user types, namely:

- **Socialisers:** Motivated by want to interact with others and create social connections.
- **Free spirits:** Motivated by autonomy and self-expression since they want to create and explore
- **Achievers:** Motivated by Mastery and are looking to learn new things, improve and challenge themselves
- **Philanthropists:** Motivated by purpose and meaning. Altruistic group looking to give and enrich with no expectation of reward.
- **Players:** Motivated by rewards. In it for themselves to collect rewards from the system.

- **Disruptors:** Motivated by change. Want to disrupt the system directly or through others to force positive or negative change.

Users can usually fit into one of these categories and in some instances they can evolve from one type to another.

4.2.c Quantitative analysis

The quantitative analysis employed a multi-stage approach to characterize curator engagement, implemented utilizing standard R[31] packages. The study relied exclusively on raw activity reports submitted by participating scientific resources, treating each submission as a verified instance of value creation as defined by the resource's own curation standards, without inferential pre-processing or cross-resource standardization. The methodology proceeded in three phases: first, a descriptive assessment of activity distributions and monthly trends was conducted to map baseline temporal behaviors and event granularity. Second, to evaluate the specific impact of the APICURON integration, the prophet forecasting library was utilized to model pre- and post-integration trends, supported by changepoint analysis to detect structural shifts in curation volume. Finally, to capture non-linear dynamics and oscillatory patterns, a Continuous Wavelet Transform (CWT) was applied to generate time-frequency scalograms. The resulting feature matrices were normalized and processed via Principal Component Analysis (PCA) to extract latent factors, which informed a Partition Around Medoids (PAM) clustering approach to distinguish specific curator behavioral profiles.

4.2.d Qualitative Analysis

The qualitative approach included feedback and input from curators which was useful for identifying key issues with implementations and also measured user attitudes towards current gamification implementations as well as identifying the user type. The two main surveys used were:

- Survey I: General attitudes towards crediting and current APICURON Implementations
- Survey II: curator motivator identification.

Survey 1 is aimed at identifying the sentiment curators presented towards crediting and recognition mechanisms. Specifically the mechanisms APICURON offers. While Survey 2 is a standardized survey from Andrzej Marczewski's user types hexad[39] to identify motivators within the curator profiles

4.3 Results

The implementation study integrated APICURON across a diverse cohort of biological databases. This yielded usage data that provided the statistics underpinning the effectiveness of crediting and its impact.

4.3.a Qualitative Analysis

From the qualitative analysis; in Survey I on attitudes the emerging typical curator profile is a professional middle aged academic who is not interested in digital reward mechanisms. The primary group that was sampled indicated that their participation in curation was not affected by the currently implemented crediting strategies. Survey “2” on the user Motivation showed that the curators who participated fall into the “Philanthropist” and “Achiever” user types but also showed signs of evolving into “Free Spirit” and “Socialiser” types.

Statement	Yes	No	Undecided
Badges matter to me	24.9%	57.2%	17.9%
Medals matter to me	25%	50%	25%
My position on the leaderboard matters to me	28.5%	42.9%	28.6%
My participation is affected by badges, medals and the leaderboard	25%	60.7%	14.3%

Table 4.1: Survey I responses to the importance of the implemented crediting elements Badges, Medals and Leaderboard

It is important to note that the survey results are from a small sample of the curators who responded to the survey. Only 28/83 (34%) of the curators participated in Survey I and 8/83 (10%) in Survey II. Owing to the low curator participation, these results on their own are therefore indicative and offer a limited insight but they do support the observations made for the quantitative analysis and suggest that the current crediting implementations have had little to no effect on curator behavior with regards to the number of frequency of curations. Even though the leaderboard had the least negative response of the 3 crediting implementations, the results still show that the majority of the curators are indifferent to the implementation of the crediting mechanisms.

4.3.b Quantitative Analysis

General statistics from the quantitative exploratory analysis revealed that the database’s total number of reports were ranked in the following decreasing or-

der: DisProt[27], Reactome[22], Pfam[13], Intact[9], Complex Portal[3], BioModels[19].

The residual and seasonality analysis of the combined monthly report distributions showed that most of the databases had their greatest number of reports during the third quarter of the year (Q3: August—October). However the data contained complex seasonality patterns and had multiple significant residuals when the lagged features to capture relationships with past values and rolling statistics were analysed; indicating various temporal dependencies within the data that could not be attributed to any one particular trigger. This was also observed when HMM models were used to measure if certain known events such as Conference paper submission and journal publication dates as well as the type of curation topic had any effect on the number of reports generated. The results suggested that the type of curation activity had no role but the journal date of a relevant journal to particular databases had somewhat of an effect on the number of curations around the time.

From fig. 4.1 the total number of contribution report distributions is shown. It can be noted that DisProt and Reactome have the greatest number of reports: 58411 and 28248, respectively. Conversely, Complex Portal and BioModels have the least amount: 6264 and 982 respectively. The gap in activity is large and DisProt has more reports than all the other resources combined due to the high granularity.

The Prophet forecast with trend change points performed across the entire dataset illustrated in fig. 4.3. Shows a consistent activity prior to the integration of APICURON. With a slight increase in activity after the implementation study. Although these fluctuations show a slight increase in growth and trigger 3 changepoints and do occur after the inclusion of APICURON. This increase lacks significance to be attributed to APICURON.

The Prophet decomposition of temporal components fig. 4.4 revealed a consistent upward trend in curation volume over the study period, accompanied by distinct weekly seasonality. Specifically, the analysis highlighted a pronounced disparity between weekday and weekend activity, with contributions peaking mid-week and dropping precipitously on Saturdays and Sundays. This cyclical pattern strongly suggests that curation is primarily performed within the constraints of professional employment schedules, distinguishing it from the irregular patterns often associated with volunteer or leisure-based crowdsourcing. Furthermore, annual seasonality metrics identified a significant surge in report volume during the late third to fourth quarter (August—October).

The Continuous Wavelet Transform used on the entire dataset, produced a time-frequency map known as a scalogram which shows the strength of a par-

ticular wavelet transform at a particular point. It shows the signal's frequency content across different time points and is valuable for understanding how the signal components evolve over time, revealing oscillatory patterns, trends and dynamics in the data. The resulting matrix from the scalogram was then normalised, correlated and then Principal Component Analysis was used to determine relevant factors (vectors of components) to be used for the PAM clustering which produced fig. 4.5.

The clustering shows two main groups with the majority of the curators in cluster 1 and the minority in the teal cluster. The main takeaway from the clustering was based on the pattern of curation and not necessarily amount. An example of this is illustrated in ??.

At certain points in time some curators exhibited similar curation patterns. However, the drivers (e.g: time of year, work, personal, etc.) behind this are not necessarily known and a study of this would reveal/give a better understanding of the impact of crediting mechanisms but more information is still required. Further analysis using hierarchical clustering may be used for this along with other surveys focused on profiling curators.

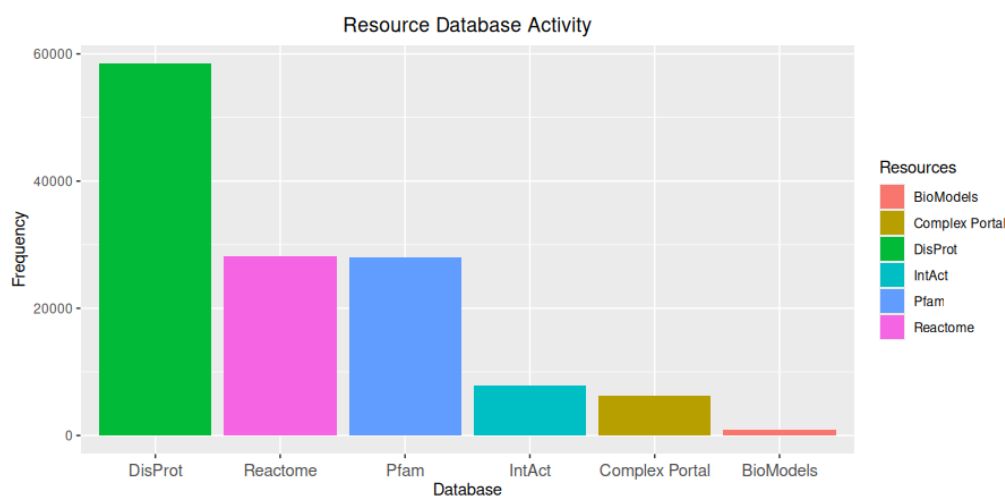


Figure 4.1: Contribution reports distribution by resource in 2025, showing the number of reported activities per resource were ranked in the following decreasing order: DisProt, Reactome, Pfam, Intact, Complex Portal, BioModels

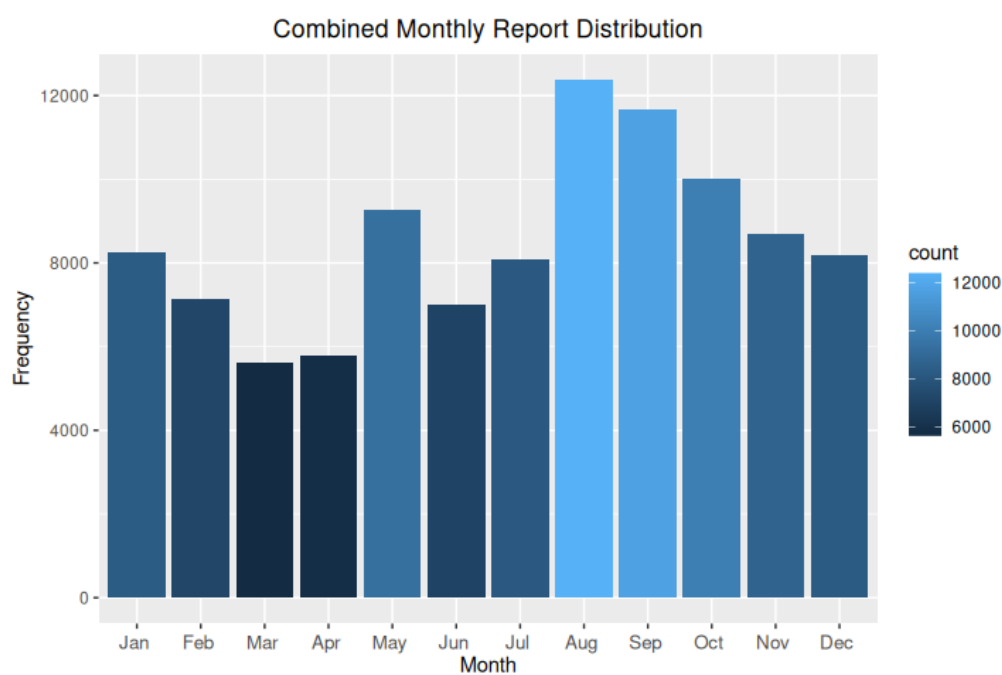


Figure 4.2: Monthly distribution of submitted contribution reports for all of the participating scientific resources.

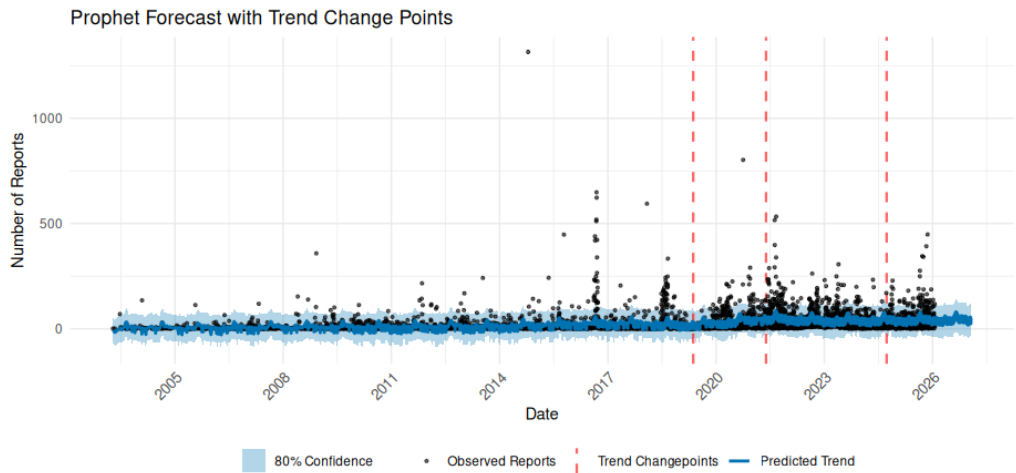


Figure 4.3: Prophet forecast with trend change points showing the overall trend of curation activity across all resources over time. The vertical dashed lines indicate detected changepoints in the trend.

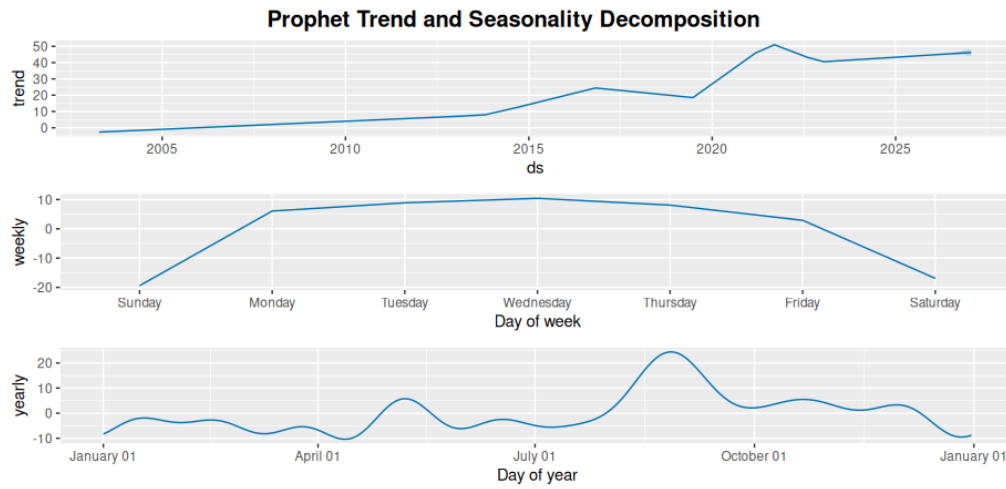


Figure 4.4: Prophet decomposition of temporal components showing the trend, weekly and yearly seasonality of curation activity across all resources over time.

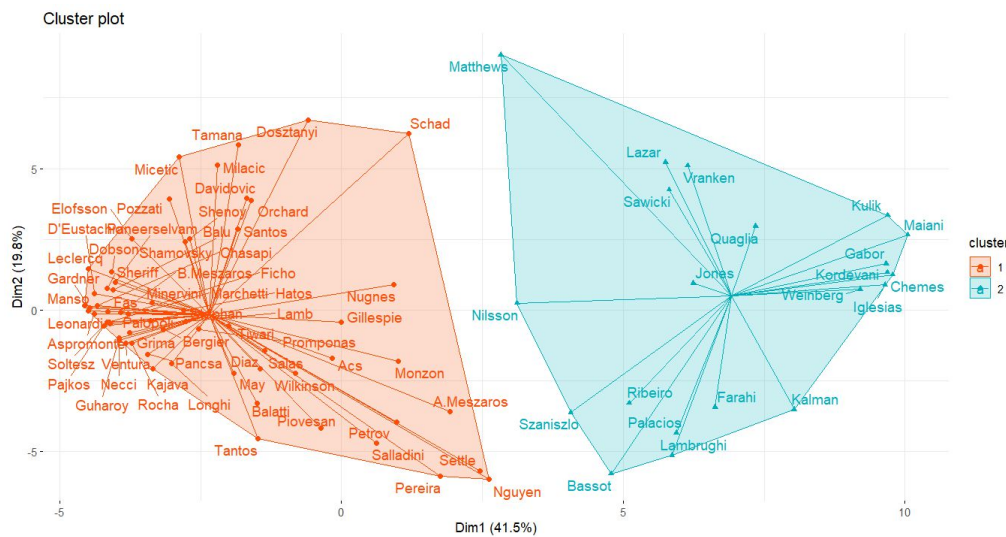


Figure 4.5: Curator Clustering using PAM showing two main clusters based on curation patterns over time rather than amount.

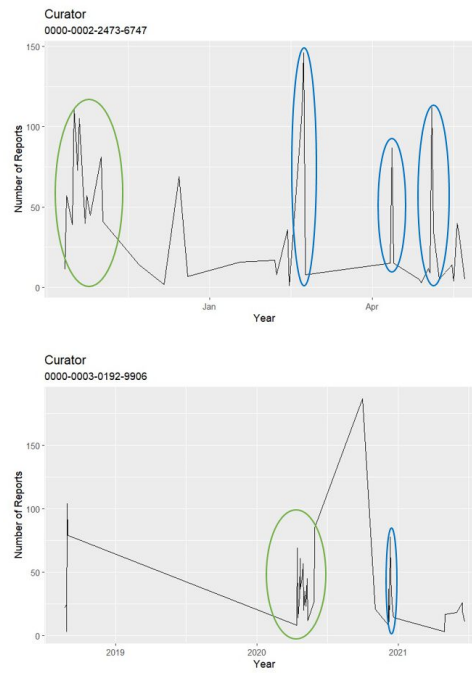


Figure 4.6: Example of two curators from different resources exhibiting similar curation patterns over time despite differing total contribution volumes.

4.3.c DisProt

DisProt is a database dedicated to the manual curation of Intrinsically Disordered Proteins. Curators at DisProt review literature to identify and annotate regions of intrinsic disorder, focusing on capturing experimental evidence and functional relevance.

With DisProt's focus on increasing annotation quality, the recognition model adopted by the platform yielded a high granularity of events, resulting in the largest number of events captured in the platform and having the biggest number of event types. [fig. 4.7](#) shows the types of activities defined by DisProt ranked by their reported occurrence, showing a high disparity between critical activities required for a valid annotation and optional activities improving the quality of the annotations

The DisProt prophet forecast results, [figs. 4.8](#) and [4.9](#) show a slightly decreasing trend through the first 3 to 4 years. Highlighting the increased difficulty in finding novel data throughout time, alongside the increased friction in improving annotation quality. It is interesting to note that after 2020/2021 the gradient of the decreasing trend becomes less steep and more stable throughout the years. The trend & seasonality analysis confirms these findings, highlighting in-addition the overall declining trend. In addition, the seasonality analysis shows the disparity between weekday and weekend activity indicating that curation is primarily performed within the constraints of professional employment schedules.

Additional prophet forecast analysis has been conducted on the dataset, specifically before the introduction of APICURON and the reward mechanisms (pre-crediting) and after the introduction of APICURON and crediting to DisProt (post-crediting) [fig. 4.10](#). The Prophet forecasts illustrate a slight shift in user behavior following the integration of APICURON. The Pre-crediting phase is characterized by stochastic engagement interspersed with significant periods of dormancy. In contrast, the Post-crediting phase shows a stabilization of the workflow. The contribution model is converted from sporadic to continuous. While extreme daily peaks were reduced, the system established a reliable baseline of recurring activity, significantly lowering variance and allowing the model to project a stable, predictable trend.

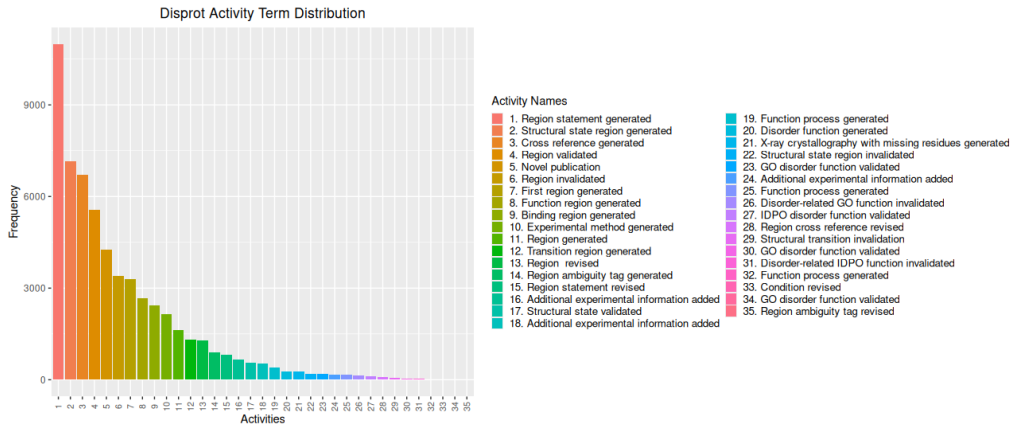


Figure 4.7: DisProt activity types distribution showing the number of reported activities per type ranked in decreasing order

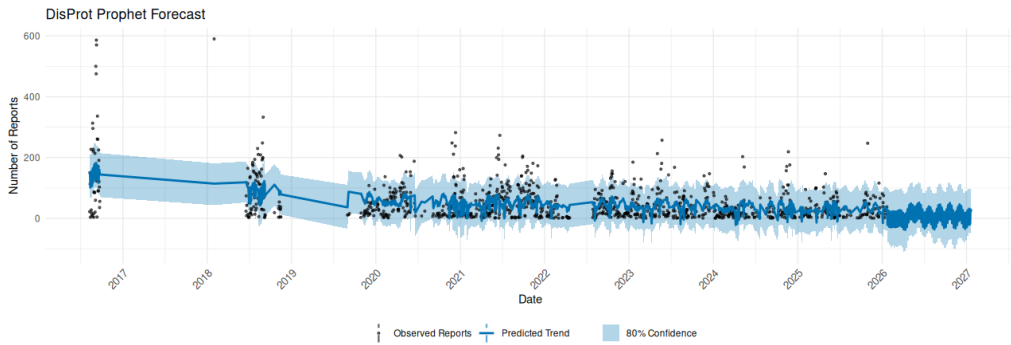


Figure 4.8: DisProt Prophet Forecast, showing a slightly decreasing trend throughout its initial years followed by a consistent trend post 2020

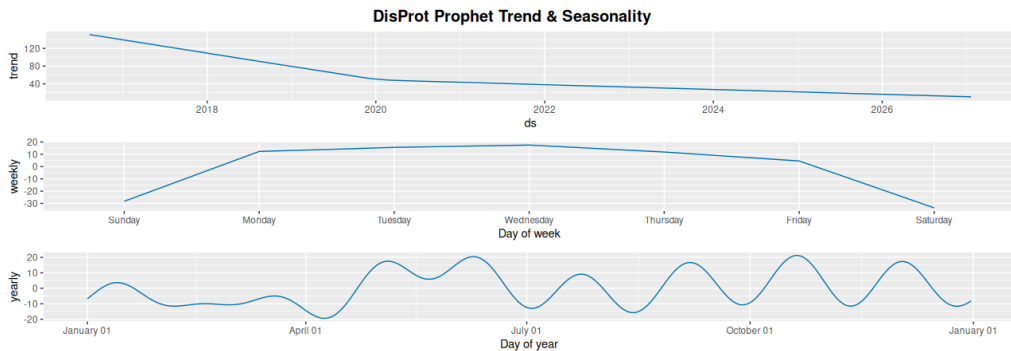


Figure 4.9: The DisProt Prophet Trend & Seasonality Analysis showing the overall declining trend, the disparity between weekday and weekend activity and an overall fluctuation of activity throughout the year, with peaks of contributions around Q2 and low activity on Q1



Figure 4.10: The DisProt Prophet Forecast showing the pre-crediting and post-crediting phases a reduced variance with lower peak volumes but a more consistent higher base-line, yielding improved predictability.

4.3.d Reactome

Reactome[22] is an open-source, open access, manually curated, and peer-reviewed pathway database focused on human biological pathways and processes. It provides intuitive bioinformatics tools for the visualization and analysis of pathway knowledge to support clinical research.

Reactome’s logging mechanism distinguishes effectively between varying types of curation efforts. fig. 4.11 illustrates this distribution, separating primary curation activities from review and maintenance contributions. This distinction is critical for the recognition framework, as it successfully captures and visualizes the often invisible maintenance work, essential for preserving data quality but frequently overlooked

The temporal analysis reveals a structured professional workflow within the Reactome community. The seasonality analysis fig. 4.13 shows a clear disparity between weekday and weekend activity, aligning with the findings in the general quantitative analysis that curation is treated primarily as a professional, work-hours activity.

Additionally, the monthly activity distribution fig. 4.12 identifies a robust cyclical pattern with recurrent peaks in February, May, August, and November. These quarterly surges are consistent throughout the dataset and likely correspond to the platform’s quarterly release cycles, suggesting that community activity is strongly driven by external operational deadlines.

The longitudinal analysis over the 22-year period reveals a shift in the platform’s growth trajectory. As shown in the trend component of the analysis fig. 4.13, Reactome experienced consistent growth until approximately 2015, after which the trend transitions into a steady decline in activity volume leading up to 2025. Despite this reduction in volume, the Prophet forecast fig. 4.14 indicates that the baseline of activity remains stable, albeit with declining peak intensities.

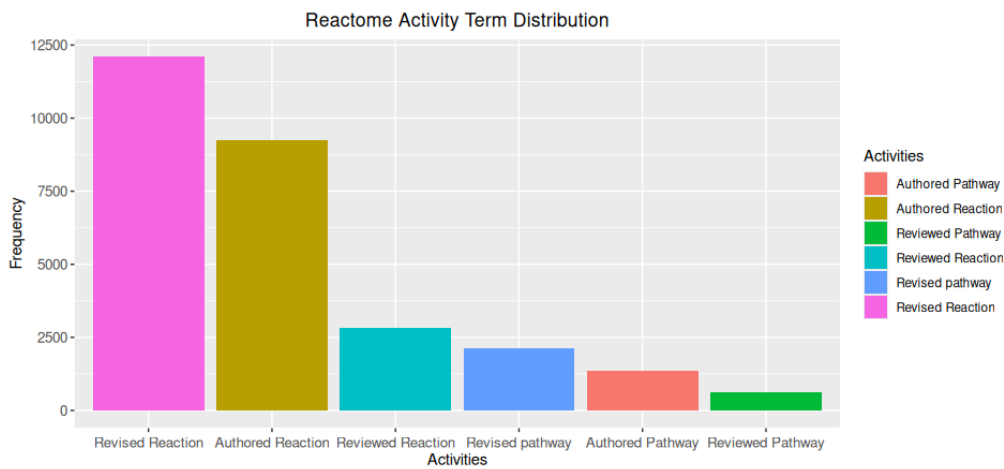


Figure 4.11: Reactome activity types distribution showing the number of reported activities per type including review and maintenance contributions ranked in decreasing order

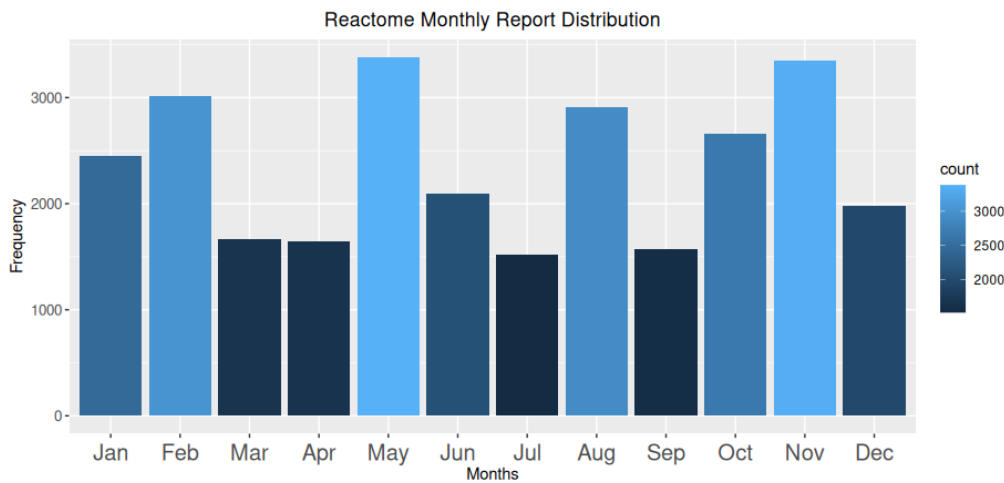


Figure 4.12: Monthly distribution of submitted contribution reports for Reactome from 2003 to 2025. Showing peak activity during May, November, February and August in order throughout the year

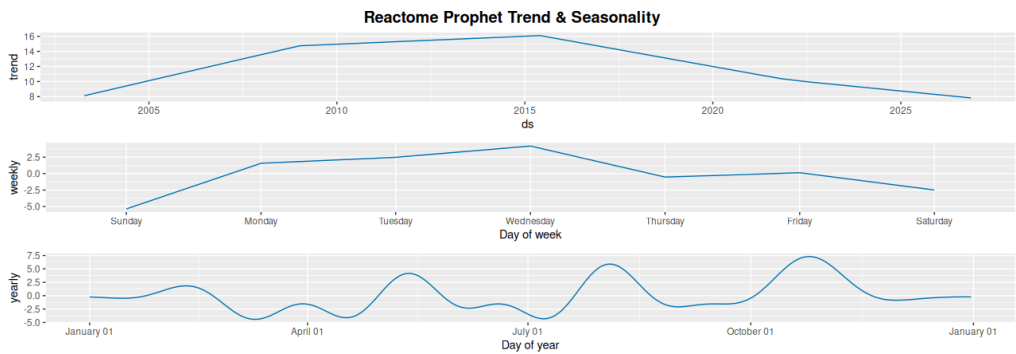


Figure 4.13: The Reactome Prophet Trend & Seasonality Analysis showing the overall trend, the disparity between weekday and weekend activity and a consistent yearly activity with peaks in February, May, August and November

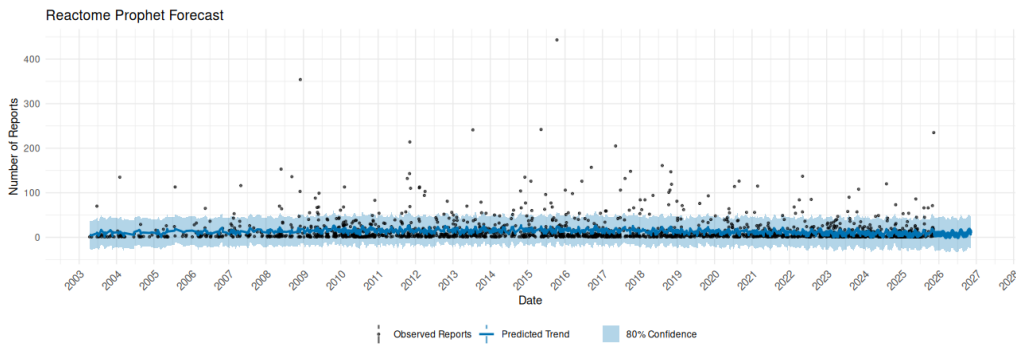


Figure 4.14: Reactome prophet forecast analysis. Showing a consistent trend with declining peaks throughout the entire 22 year period.

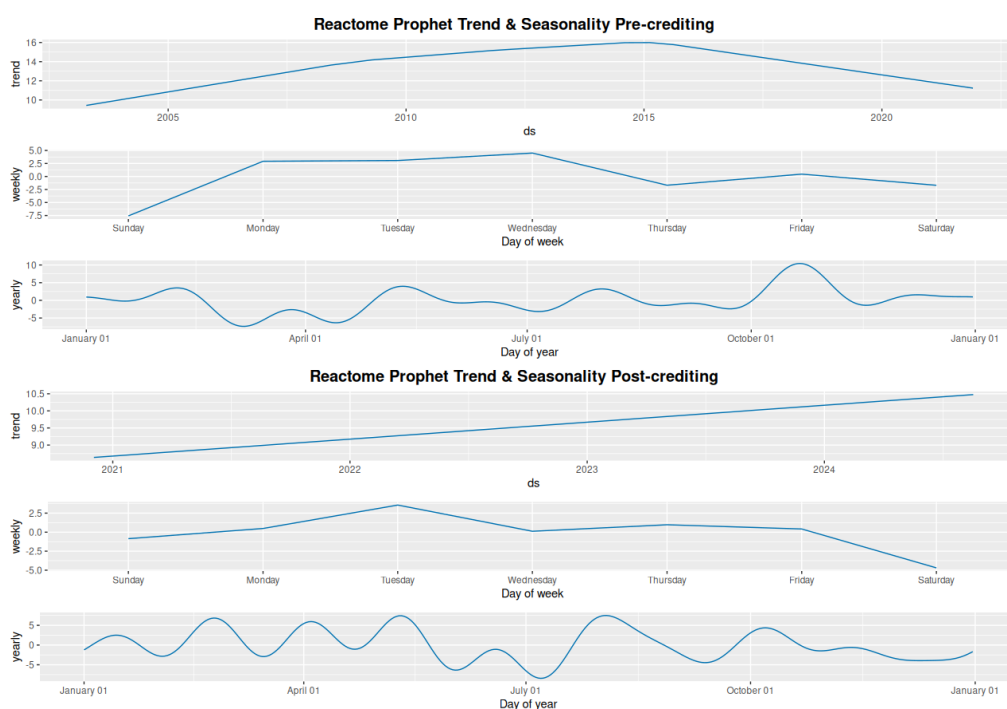


Figure 4.15: Reactome prophet trend & seasonality pre/post crediting shows a low volume increase in overall trend in post-crediting phase opposing a decline at the end of the pre-crediting period. And a similar and consistent seasonality across the dataset

4.3.e Pfam

Pfam[13] is a significant resource comprising a large collection of protein families, each defined by multiple sequence alignments and hidden Markov models (HMMs). The biocuration activity at Pfam centres on defining new protein families, maintaining the accuracy of existing family boundaries, and generating sequence alignments and HMMs.

The integration of Pfam data into the APICURON ecosystem revealed specific limitations regarding metadata resolution. As illustrated in fig. 4.16, the credit attribution model inherited from Pfam’s dataset is notably coarse-grained. The vast majority of contributions fall under the generic umbrella term “Updating a family,” which obscures the specific nature of individual contributions. This lack of categorical granularity suggests that while the volume of work is captured, the specific type of labor remains difficult to distinguish without upstream changes to the tracking mechanism

Despite the categorical limitations, the temporal quality of the dataset is robust, offering a consistent view of community engagement over time. The temporal analysis reveals a distinct seasonal pattern, characterized by a recurrent peak in activity during August fig. 4.17. This observation is further corroborated by the Prophet trend and seasonality analysis shown in fig. 4.20, which highlights a generally low and stable baseline of activity throughout the year, punctuated by these specific seasonal spikes. Additionally, the seasonality analysis highlights a clear disparity between weekday and weekend activity, aligning with standard professional engagement patterns observed in the general quantitative analysis.

Analysis using the Prophet forecasting model fig. 4.18 demonstrates a steady baseline increase in contributions throughout the platform’s lifetime. The model captures the recurrent yearly spikes—likely driven by annual curation events—and projects a significant surge in activity for the year 2025.

To assess the impact of the crediting mechanism, the forecasting analysis was segmented into pre- and post-crediting phases fig. 4.19. This comparison aligns with the initial findings, showing a stable, steadily increasing trend across the entire dataset. This consistency suggests that the community’s growth trajectory is stable, with the projected increase in 2025 following the established historical momentum rather than a deviation caused by the crediting intervention alone.

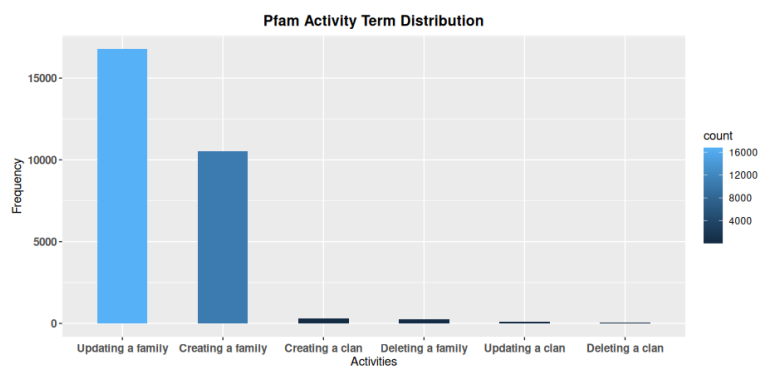


Figure 4.16: Pfam activity types distribution showing the number of reported activities per type ranked in decreasing order and highlighting the majority of reported activities falling under the umbrella term "Updating a family"

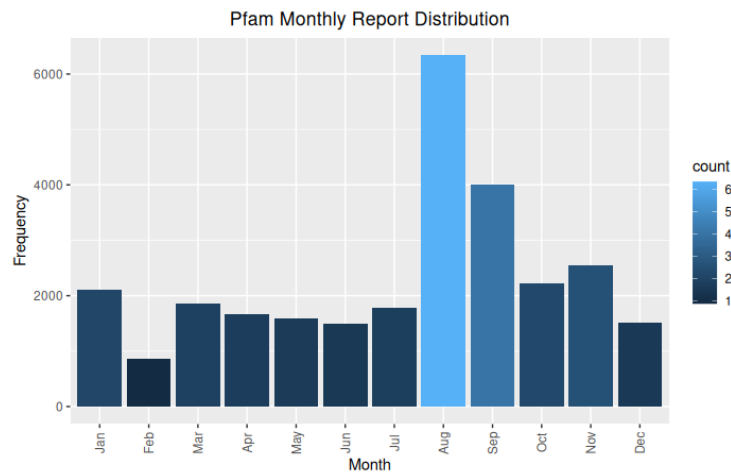


Figure 4.17: Monthly distribution of submitted contribution reports for Pfam, showing peak activity during August

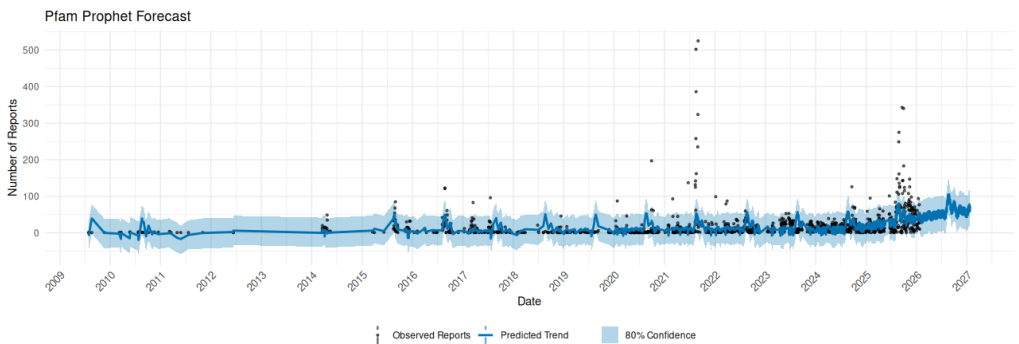


Figure 4.18: Pfam prophet forecast analysis showing a steady baseline increase in contributions throughout the platform’s lifetime with recurrent yearly spikes and a projected significant surge in activity for the year 2025

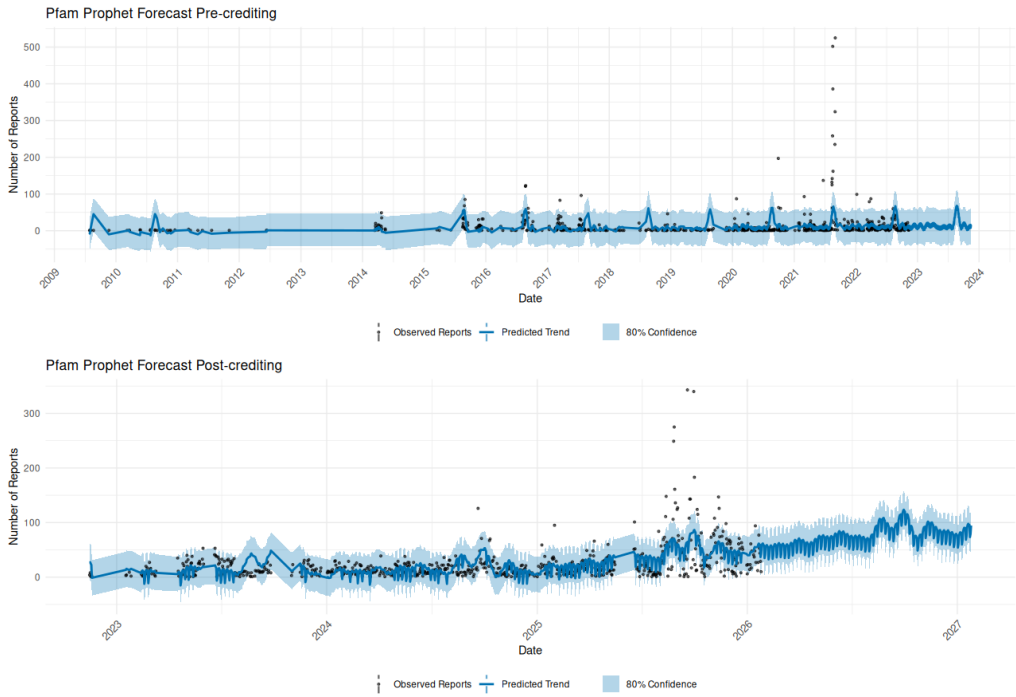


Figure 4.19: Pfam prophet forecast showing the pre-crediting and post-crediting phases with a stable, steadily increasing trend across the entire dataset

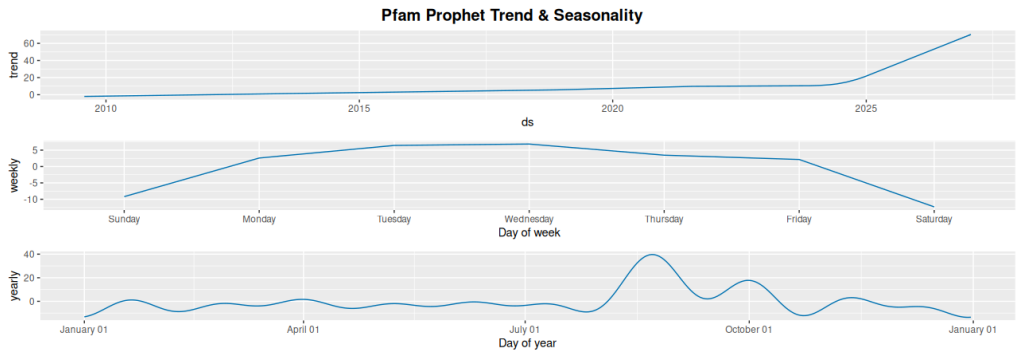


Figure 4.20: The Pfam Prophet Trend & Seasonality Analysis showing an overall stable trend, the disparity between weekday and weekend activity and a consistent yearly activity with peaks in August

4.3.f IntAct

IntAct[9] is an open-source database that makes protein-protein interaction data derived from the scientific literature freely available. Interaction curators read publications, extract interaction details such as interacting partners, experimental method, and supporting evidence, to subsequently map this information to standard ontologies. APICURON plays a simple role for Intact, ensuring contributions are credited when publications are curated and reviewed

IntAct defines only 2 types of activities, making the recognition data obscure and difficult to interpret. This has been mainly due to GDPR compliance concerns regarding user data. Which hinders the platform’s ability to credit for often overlooked curation work.

The temporal analysis of IntAct curator activity reveals a distinct seasonal pattern characterized by surged engagement during the winter months. As illustrated in the monthly report distribution fig. 4.21, activity peaks notably in December and January, a recurrence further corroborated by the trend and seasonality decomposition fig. 4.22. Despite these seasonal fluctuations, the longitudinal Prophet forecast analysisfig. 4.23 demonstrates a remarkably consistent and stable baseline of activity throughout the entire dataset. However, this stability is punctuated by a significant, singular surge in activity observed in 2014, representing a deviation from the otherwise steady curation trend.

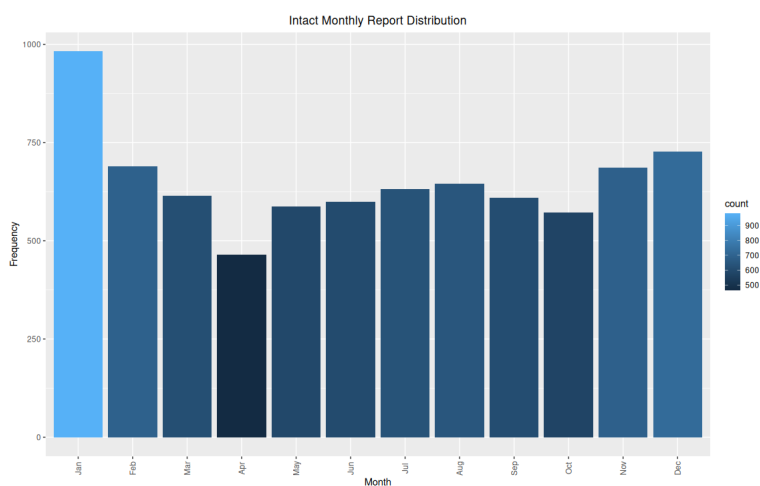


Figure 4.21: Monthly distribution of submitted contribution reports for IntAct, showing peak activity during December and January

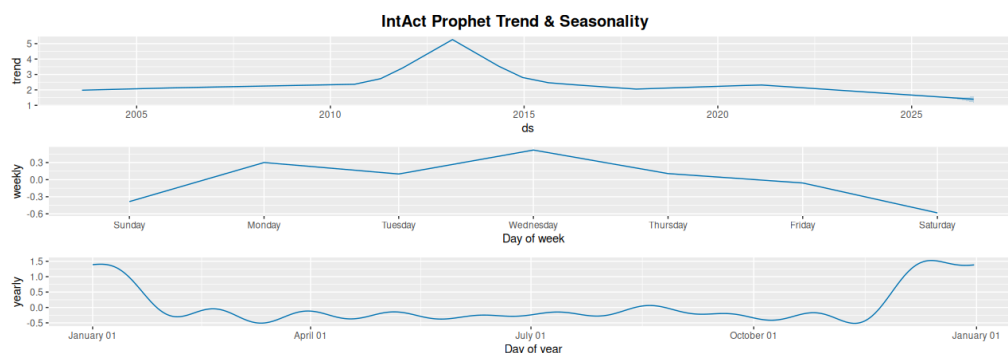


Figure 4.22: The IntAct Prophet Trend & Seasonality Analysis showing the overall stable trend, the disparity between weekday and weekend activity and a consistent yearly activity with peaks in December and January

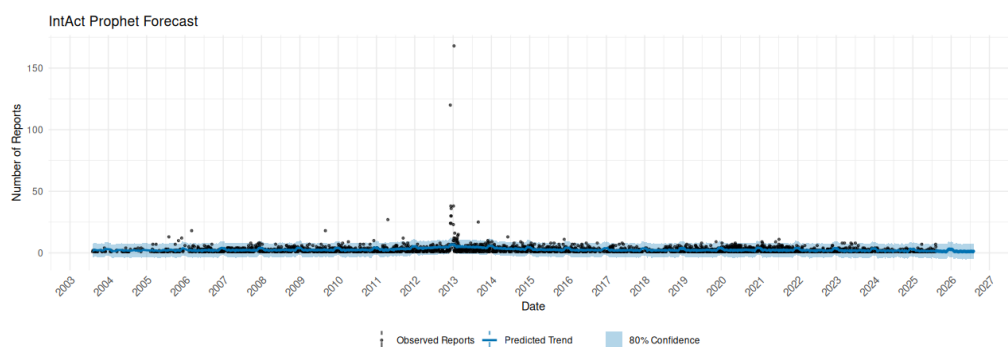


Figure 4.23: IntAct prophet forecast analysis showing a stable baseline of activity throughout the entire dataset with a significant surge in activity observed in 2014

4.3.g Complex Portal

Complex Portal is a manually curated, encyclopaedic resource of stable macromolecular complexes (proteins and/or nucleic acids) from various organisms, supplemented by machine-learning predicted human complexes. Curation defines stoichiometry, components, and functional context based on experimental evidence and adheres to specific guidelines regarding complex type and size. APICURON is utilized for crediting complex curation and review.

For the APICURON integration, Complex Portal adopted a minimal recognition model, defining only two distinct activity types: Complex Curation and Complex Approval. Similar to the IntAct use case, this limitation was driven primarily by internal interpretations of GDPR compliance, which restricted the transmission of granular user data.

Consequently, while APICURON successfully captured the volume of activity, the semantic richness of the curation workflow was lost, rendering the resulting credit less descriptive than in the DisProt pilot. Despite these semantic limitations, the quantitative signals captured by APICURON reveal the following insights:

- The Prophet forecast fig. 4.25 indicates a positive trajectory with sustained growth in curation activity over time. The data shows a pattern of increased volatility and higher contribution volumes following 2020.
- The trend analysis fig. 4.26 reveals a distinct disparity between weekday and weekend activity, with contributions dropping significantly on weekends. This suggests a professionalized curation model where contributors engage primarily during standard working hours.
- The seasonal trend analysis fig. 4.26 identifies a recurring underlying increase in activity around August. However, the raw monthly distribution fig. 4.24 displays a dominant peak in October. This indicates that while late summer is a trend for steady activity, the absolute volume of the database has been driven by specific, high-intensity submission events occurring in October.

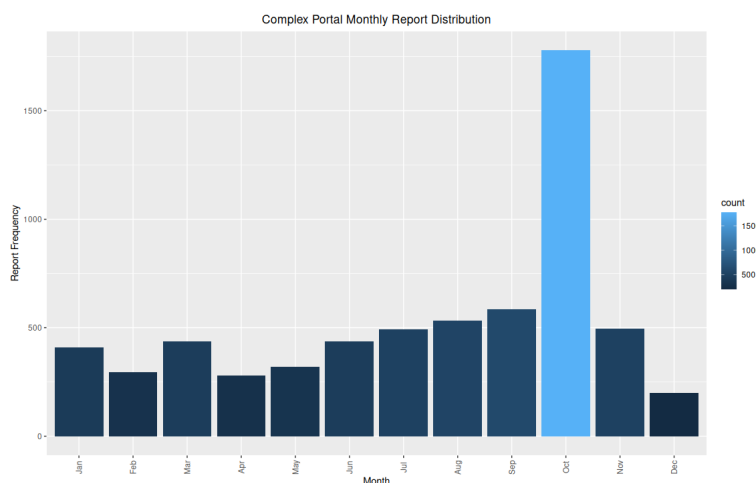


Figure 4.24: Monthly distribution of submitted contribution reports for Complex Portal, showing peak activity during October

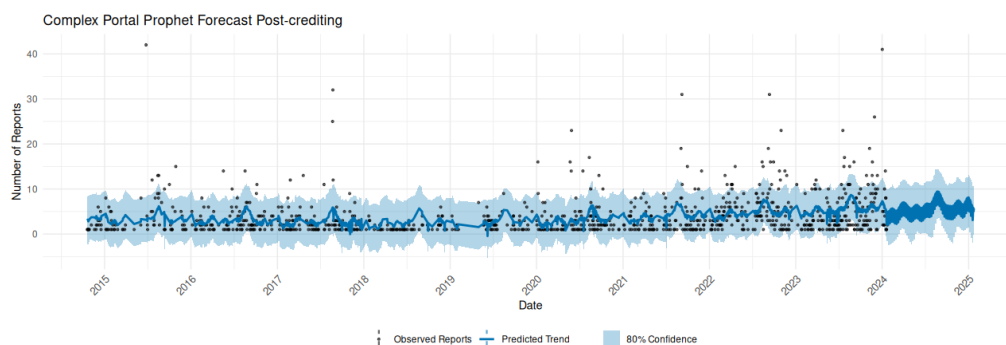


Figure 4.25: Complex Portal prophet forecast analysis showing a positive trajectory with sustained growth in curation activity over time, with increased volatility and higher contribution volumes following 2020

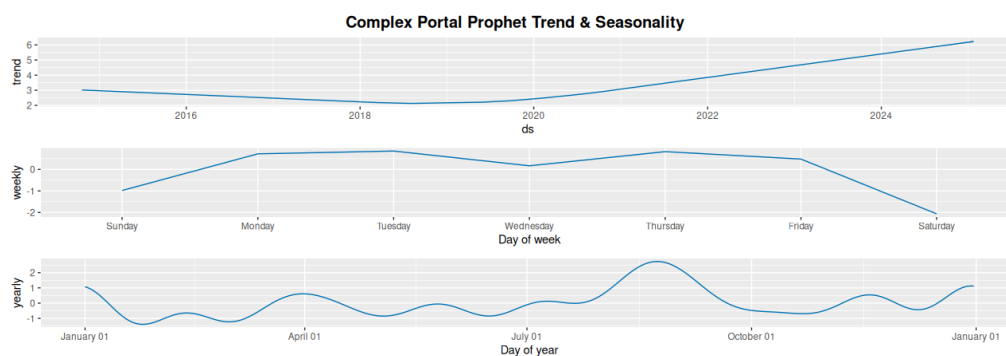


Figure 4.26: The Complex Portal Prophet Trend & Seasonality Analysis showing the overall increasing trend, the disparity between weekday and weekend activity and a consistent yearly activity with peaks in August

4.3.h BioModels

BioModels[19] is a repository of mathematical models representing diverse biological systems, with models sourced from peer-reviewed literature and curated to be semantically enriched and mathematically correct. As an early adopter, APICURON is leveraged by BioModels to recognize and credit contributions to mathematical models from initial submission and curation/annotation to the creation of new versions. This encourages curators to submit, enhance and maintain mathematical models on the BioModels repository.

The temporal data fig. 4.28 shows a stable baseline punctuated by sporadic spikes in volume. The overall decreasing trend observed in Figure 29 suggests the resource has moved from an initial high-growth phase into a steady-state maintenance mode. The visible gap in 2022 represents a temporary operational hiatus in data ingestion. Activity is strongly coupled with the academic calendar

figs. 4.27 and 4.29. There are significant peaks in Q1–Q2 and Q4, with a notable drop during the summer months (Q3). The sharp disparity between weekday and weekend activity fig. 4.29 confirms that curation is performed almost exclusively during professional working hours.

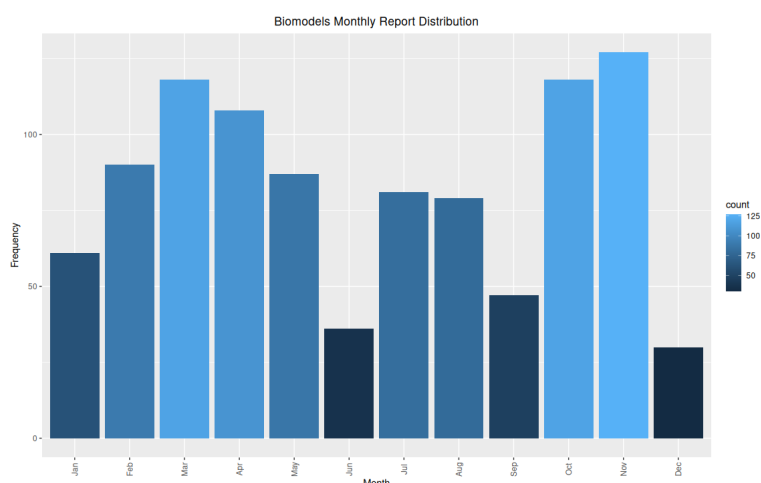


Figure 4.27: Monthly distribution of submitted contribution reports for BioModels, showing peak activity during Q1–Q2 and Q4 with a notable drop during the summer months (Q3)

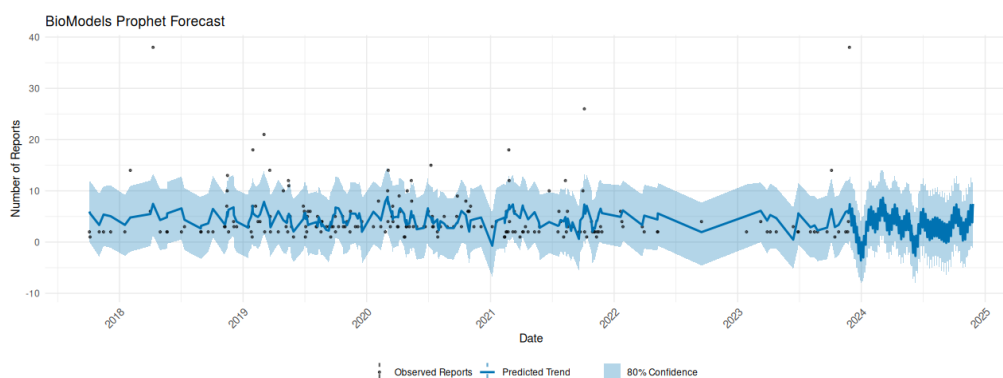


Figure 4.28: BioModels prophet forecast analysis. Showing stable activity throughout the entire period with sporadic spikes in activity and a gap of inactivity in 2022.

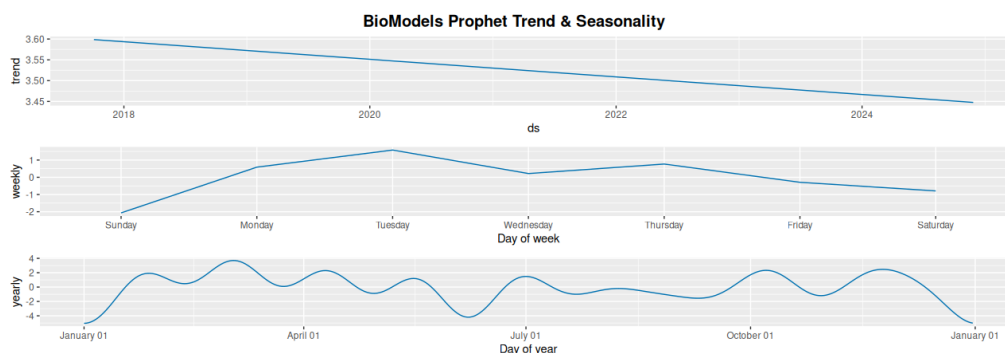


Figure 4.29: The BioModels Prophet Trend & Seasonality Analysis showing the overall decreasing trend, the disparity between weekday and weekend activity and a consistent yearly activity with peaks in Q1-Q2 and Q4 and a drop in Q3

4.4 Implementation & Analysis Outcomes

The deployment of APICURON across this heterogeneous cohort provided critical empirical data regarding the feasibility of a universal credit attribution system. Beyond the successful technical integration of the partner resources, the study yielded fundamental findings that have since driven the architectural and conceptual evolution of the platform.

4.4.a Reports on Biocuration Activities

The contribution data obtained during the implementation study represented an opportunity to analyze behavior trends in biocuration. Additionally the implementation helped to identify the impact of acknowledgement on biocurators.

4.4.b Validated Data Model

The study demonstrated that the APICURON data model possesses the necessary semantic flexibility to capture this broad spectrum of scientific contributions without fragmentation. By abstracting contributions into a unified event schema, the system successfully harmonized varying contribution workflows under a single representation. This finding confirms the feasibility of a universal credit standard, proving that a single “currency” of recognition can exist across the fragmented biocuration ecosystem.

4.4.c Sociotechnical Findings

A critical qualitative finding of the study was the varying reception of public ranking mechanics across different curator cohorts. While community volun-

teers often responded positively to crediting, feedback from professional curators at established resources indicated a distinct discomfort with being ranked.

This friction highlighted that for professional staff, the Leaderboard mechanic was perceived not as an incentive, but as an intrusive metric of their daily employment. This finding validated the necessity for a Privacy-First design, distinguishing between the attribution of credit and the public exhibition of activity. Consequently, this feedback drove two major feature implementations:

- **Contribution Groups**

A mechanism to segregate recognition mechanisms to provide fair comparative contexts within a scientific resource. Ensuring professionals are not ranked against community volunteers.

- **Granular Privacy Controls**

The introduction of “visibility toggles”, granting users the right to opt-out of public profiles entirely while still accumulating credit in the background. This established the “Right to Non-Display” as a fundamental tenet of the platform.

4.4.d The Consent Gap

The most significant outcome of the implementation study was the identification of a critical regulatory gap. While the technical connection between APICURON and ORCID was established, the legal propagation of user data raised immediate GDPR compliance issues .

The partner resources lacked a standardized mechanism to capture and transmit explicit user consent for this specific data processing purpose. The study demonstrated that technical interoperability is insufficient without regulatory interoperability; the system could technically push data to ORCID, but legally, it required a verifiable audit trail of user permission. This Compliance Gap could not be solved by code alone; it required a dedicated infrastructure for consent management. This discovery serves as the primary motivation for the Consent Management Platform detailed in the subsequent chapter.

4.4.e Broadening scope

Finally, the implementation study revealed that the core entities of the APICURON schema are not intrinsically bound to the domain of biocuration. By successfully mapping the distinct data models of the partner cohort (e.g., mathematical models in BioModels vs. structural annotations in DisProt) to a unified

event ontology, the system demonstrated a high degree of flexibility.

This finding indicates that the framework's applicability extends beyond biological metadata. The system's architecture is theoretically capable of tracking contributions in various research domains involving granular digital interactions, such as open-source code contributions or the development of educational training materials. This realization has expanded the project's long-term scope, validating APICURON not merely as a biocuration tool, but as a generalized infrastructure for crediting diverse research outputs.

4.5 Conclusion

This implementation study successfully validated the APICURON framework across a representative cross-section of the biocuration ecosystem. It demonstrated that a unified schema can effectively capture the Broad Spectrum of biocuration activities from the manual precision of DisProt to the automated scale of Pfam.

However, technical aggregation is only the first step. The friction encountered regarding curator privacy and the indifference toward gamification highlights a critical reality: a sustainable credit infrastructure requires as much rigor in its legal and social engineering as it does in its software architecture.

As argued in the Credit & Recognition chapter, a shift in research culture cannot be achieved through infrastructure alone. For crediting mechanisms to have substantial impact, they must retain tangible value within the research community. Therefore, while it is reasonable not to expect a singular study to reverse macroscopic trends, this work establishes the essential intermediary logic layer, proving that while value is socially determined, the capability to track it is now engineered.

Part V

The Consent Management System

The Consent Management System

The APICURON implementation study identified privacy and General Data Protection Regulation compliance as significant issues and blockers. APICURON is committed to assisting and facilitating compliance with the General Data Protection Regulation (GDPR). This document outlines a consent management system that the scientific resources leveraging APICURON could use to collect and manage consent to share personal data with APICURON and benefit from the features APICURON offers.

This section details the APICURON Consent Management System, a solution designed to streamline GDPR compliance for scientific resources, enabling lawful data sharing with APICURON and unlocking its full feature set.

5.1 Context

During the APICURON implementation study, which involved the integration setup with several European Bioinformatics Institute (EBI) resources, including IntAct[9], Reactome[22], BioModels[19], and Pfam[13], the technical implementation proceeded successfully, and all relevant resources were integrated into the APICURON platform as planned.

The study also revealed a number of privacy-related challenges, particularly in relation to compliance with the General Data Protection Regulation GDPR. In accordance with GDPR requirements, the participating EBI resources were obliged to exclude all contributions originating from users who had not explicitly provided consent for their data to be shared. Consequently, only data collected with valid and informed consent from individuals affiliated with the respective

organizations could be transmitted to APICURON.

These regulatory constraints introduced additional complexity into the data-sharing workflows and underscored the necessity of a unified, transparent, and verifiable mechanism for managing user consent across all partner resources. This experience directly motivated the present proposal, which aims to establish a dedicated Consent Management System to facilitate GDPR compliance, ensure lawful data processing, and streamline future integrations between APICURON and participating scientific resources. This experience directly motivated the present proposal, which aims to establish a dedicated Consent Management System to facilitate GDPR compliance, ensure lawful data processing, and streamline future integrations between APICURON and participating scientific resources.

5.2 Regulatory Basis

The General Data Protection Regulation is an extensive document outlining laws regarding multiple aspects of processing personal data. This section outlines the articles of the GDPR that are most relevant to the APICURON use case for current and potential partners and discusses their impact on the decisions made in this project.

5.2.a Definitions

These definitions are complete and outlined in Article 4 of the GDPR[33]

- **Person (also “data subject”):** An identifiable natural person can be identified, directly or indirectly. For this project, the person is the contributor to the resources that interact with the data controllers and has their personal data managed by them. For this project, the person is the curator and contributor on the partner resources platform.
- **Controller (also Data Controller):** is any natural or legal person, public authority, agency, or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data. For this project, the controllers are APICURON and the partner resources.

5.2.b Lawfulness of processing

Personal data processing is lawful under the GDPR only when it is linked to a valid legal basis. The primary legal basis we’re using to allow lawful data sharing

to APICURON is consent from the data subject according to Article 6 (1)(a)[33]. Therefore, the Data Subject must provide explicit informed consent to process their data. A consent they can withdraw. Moreover, the data controller needs to take procedures to handle the withdrawal of consent according to Article 7(3).

5.2.c Demonstration of Compliance

In GDPR, the Data Controller is obligated to demonstrate compliance. Processes, implementation, and mitigation strategies must be in place to prove that personal data is processed lawfully. For this project, this can include many aspects of the processing of personal data. Some examples are listed below:

- Processes, Configuration, and standards that guarantee any personal data is entirely removed upon the data subject's request according to the "right to be forgotten" (except for data that is used to demonstrate compliance. . . etc, e.g, consent-related data) Article 17 (1)(b)[33].
- Mechanisms mitigating the risk of tampering with personal data. These mechanisms should apply reasonable security measures to prevent tampering Article 5 (1)(f)[33]

5.2.d Transparency

One of the main principles of GDPR law is transparency. Articles 13 & 14 of the regulation[33] explicitly state the information to be provided to the data subject when personal data is collected, whether directly from the user or indirectly through another source. The disclosure of the legal basis for processing, the duration of the processing, and information regarding consent withdrawal or claims, etc, is all mandatory.

5.3 System Overview

This project aims to offer a solution that can help partner platforms and APICURON maintain compliance with GDPR regulations, streamlining compliance and reducing risk factors when sharing personal data. The APICURON Consent Management Platform is designed to be an infrastructure component deployed on the scientific resource's infrastructure, making compliance a more straightforward, default approach.

5.3.a Technical Design

In accordance with the regulation mentioned above, this section will outline the specifications that need to be aligned to offer proper, reasonable mechanisms enabling compliance.

1. Storage System
 - (a) Includes by-design auditing features to be used for demonstrating compliance
2. A user interface for providing and withdrawing consent
 - (a) Integrates a mechanism to present the privacy policy documents to users.
 - (b) Collects informed, explicit, and unambiguous consent.
3. A backend system offering an Application Programming Interface.
 - (a) Gives secure access to consent-related data
 - (b) Offers means to react to consent withdrawals (e.g, webhooks) actively
4. Portability
 - (a) This system needs to be easily portable and deployable on various infrastructure environments (on-premises, cloud environments, etc).

The aforementioned specification is instrumental in applying the principles of the GDPR and achieving regulatory compliance. However, it is not intended as a comprehensive solution for complete GDPR compliance. The following are examples of rules where this specification is inadequate to facilitate compliance.

The aforementioned specification is instrumental in applying the principles of the GDPR and achieving regulatory compliance. However, it is not intended as a comprehensive solution for complete GDPR compliance. The following are examples of rules where this specification is inadequate to facilitate compliance.

- **Article 15: Right of access by the data subject**

The specification includes a system that only stores consent and does not fall within other personal data that may be stored on different infrastructure components.

- **Article 17: Right to erasure**

The specification describes only a system to store consent data, which does not fall within the right to erasure, as it is a necessary processing of data to demonstrate compliance with the regulation. Any other personal data outside the scope of the specification is subject to this regulation.

5.3.b Architecture: Components & Tools

Following the technical specification mentioned above. This section will describe the concrete decisions that are taken. The solution is only composed of three components: a web server, a database, and a web component.

Web component

The web component is a user interface component that users can use to state their consent explicitly. This component is supposed to be easily embedded in the scientific resource's web interface with little effort and configuration. When the user explicitly gives their consent, this component would send a request to the web server in this system.

5.3.c The web server

The web server acts as the core of the system. Once the web component submits consent data (either offer or withdrawal of consent), the web server stores the information in the database before propagating this information to the relevant parties.

5.3.d The Database

The choice of KurrentDB (formerly EventStoreDB) is critical as its Event Sourcing pattern inherently provides a verifiable, anti-tampering audit trail by design. This native feature directly supports the GDPR's Demonstration of Compliance obligation

The Event Sourcing pattern defines an approach to handling operations on data that's driven by a sequence of events, each of which is recorded in an append-only store. Application code raises events that imperatively describe the action taken on the object.

Although the Event Sourcing pattern was initially developed to improve scalability in microservices by handling concurrent update operations, this pattern serves the aim of this project case by offering an audit trail by design. This is necessary to comply with regulations requiring verifiable, anti-tampering records for consent storage. These guarantees are essential in this case. Although traditional relational databases can offer an audit trail, they require advanced configurations that can become difficult to maintain and reproduce in different environments or through database versions.

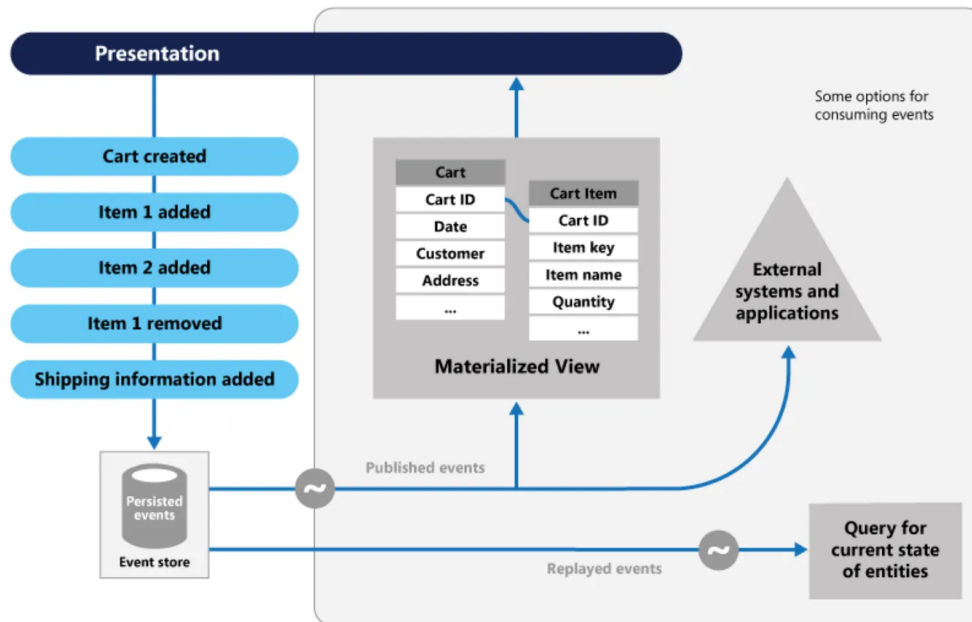


Figure 5.1: Overview of the event sourcing pattern, including some typical implementations, source:[6]

5.3.e Workflows

After identifying the elements of the system, this section will elaborate on the interactions of these elements within and outside the system. Of course, these interactions independently describe very little. Therefore, this section will describe these interactions through workflows with overlapping logical components and implementations that all serve different purposes based on the workflow.

Gather Consent Data

Using the web component, users can offer or withdraw consent. The web component would make API calls to the consent management web server, which performs multiple actions in parallel.

- Update the event log in the database with the new event
- Publish a new state in the API, which means that every subscriber on the web API can get an event about the updated consent of the user, which includes APICURON & partner resources.

Query Consent Data

The system architecture mandates that the Web Server function as the authoritative Query Interface for consent data. This interface facilitates the synchronous retrieval of the aggregated consent state for a specified data subject. Critically, this function is designed to operate independently of the system's asynchronous event-driven workflows (e.g., consent recording or update propagation). This design ensures that partner resources, utilizing this API endpoint, can consistently and immediately access the most recent, consolidated consent record necessary for lawful data processing.

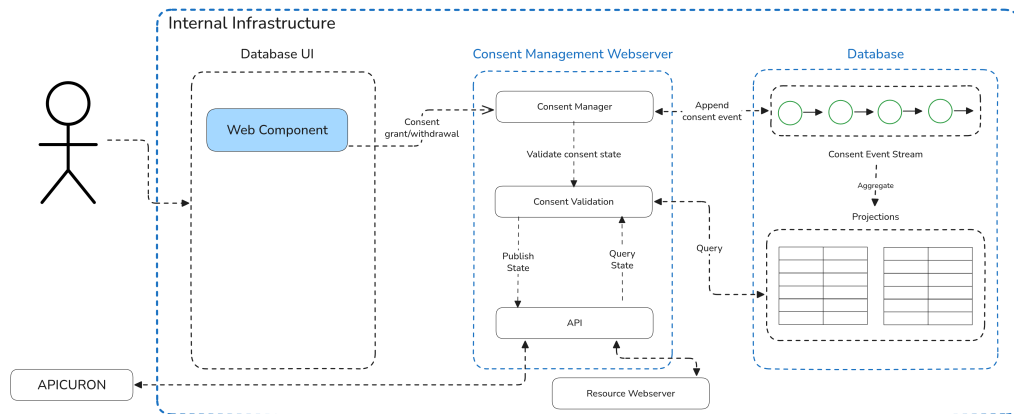


Figure 5.2: The architecture of the APICURON Consent Management Platform showing all of the interacting components, with the data subject

Part VI

**Credit & Recognition for
Software**

Credit & Recognition for Software

The technical success of the APICURON implementation study demonstrated that the model effectively captures the broad spectrum of biocuration activities. This finding indicates that the framework's applicability extends beyond biological metadata; the system's event-driven architecture is capable of tracking contributions across diverse research domains. Consequently, the logical next step for validating the APICURON model is Research Software. Like biocuration, software plays a critical role in modern research yet lacks clear, widely accepted mechanisms for credit. Furthermore, attributing credit to software is complicated by the highly fragmented and non-linear nature of modern development workflows.

The growing importance of research software and the critical role played by Research Software Engineer (RSE) in its development, maintenance, and dissemination necessitates a robust and standardized framework for their recognition. Unlike traditional academic outputs, contributions to research software often lack clear, widely accepted mechanisms for credit and attribution, leading to undervalued efforts and disincentives for community engagement. Similarly, the labor involved in developing and delivering effective training materials, often shared and iterated upon via platforms like GitHub, also goes largely unrecognized.

This chapter outlines a proposed framework for recognizing the contributions of RSEs, leveraging their shared activity on GitHub as a common ground for identifying and quantifying their impact.

6.1 The Software Attribution Paradox

As established in chapter 1, the current research assessment landscape suffers from a fundamental granularity gap. While the technical success of the API-CURON implementation study chapter 4 demonstrated that this gap can be bridged for biocuration, Research Software presents a distinct and equally critical challenge. Software has become the backbone of modern science, yet the human investment required to build and maintain it remains largely invisible within the academic record.

6.1.a Process over Product Recognition

As outlined in chapter 2. Significant community efforts have emerged to address the software recognition deficit. Initiatives such as CodeMeta and the Citation File Format have standardized the metadata necessary to treat software as a citable research output. Simultaneously, platforms like Zenodo have been instrumental in integrating software into the scholarly ecosystem by minting DOIs for specific releases, effectively allowing code to be cited like a paper. Furthermore, mechanisms like GitHub’s CODEOWNERS allow for the assignment of responsibility to specific parts of a codebase.

While these initiatives are essential for reproducibility and discovery, they primarily capture the Research Product rather than the Human Investment. They function as static snapshots: a Zenodo DOI captures a release at a single point in time, and a CODEOWNERS file defines current responsibility, not historical effort. They fail to capture the continuous, non-linear stream of intellectual labor, from bug fixes to architectural refactoring that defines modern software engineering. Consequently, a paradox persists: we can now cite the software, but we still struggle to credit the engineer. The current ecosystem conflates the tool with the labour, obscuring the granular contributions of Research Software Engineer and Trainers. To resolve this, it is necessary to move beyond crediting the static artifact and establish a framework capable of recognizing the dynamic lifecycle of software development.

To accurately capture the human investment of RSEs, reliance on manual self-reporting is insufficient as it is prone to bias and omission. Instead, recognition frameworks must tap into the provenance systems where the work actually occurs: the Version Control System (VCS).

6.1.b The Commit Fallacy

Within distributed version control systems, the commit serves as the fundamental unit of data integrity and history. While it provides a definitive technical record of code modification, the commit is intrinsically flawed when utilized as a baseline for academic credit. Commits exhibit extreme variance in granularity and impact: a single unit may represent a trivial syntax correction or a fundamental architectural refactoring.

Critically, this limitation extends beyond the commit itself to the broader ecosystem of VCS platform features. Native analytical tools provided by platforms like GitHub (e.g., contribution graphs, activity dashboards, and pulse insights) are fundamentally aggregations of these atomic units. Because these higher-level features utilize the raw commit as their underlying data structure, they inherit its semantic opacity. A count of pull requests opened or releases published is ultimately a derivative of the commit history and suffers from the same inability to distinguish between administrative maintenance and intellectual contribution. Consequently, neither the raw commit nor the native platform metrics built upon it can serve as a reliable proxy for scientific credit without a supplementary semantic layer.

6.1.c The Semantic Gap

This presents a distinct challenge for the APICURON model. In the biocuration implementation study chapter 4, the model relied on explicitly defined activities (e.g., “annotate protein”) that possessed inherent scientific value as defined by scientific resources. In contrast, software development generates a noisy stream of low-level technical events (pushes, merges, issue comments) that lack immediate scholarly meaning.

6.2 The Research Software Credit Schema

To bridge the semantic gap between raw version control events and meaningful academic recognition, this thesis introduces the Research Software Credit Schema. Conceptually, the Research Software Credit Schema (RSCS) functions as the software-specific implementation of the Activity Definition introduced in chapter 2. It acts as an interpretative middleware layer, translating the noisy stream of technical operations (e.g., commits, merges, issue comments) into standardized, high-value recognition tokens that the APICURON ecosystem can ingest.

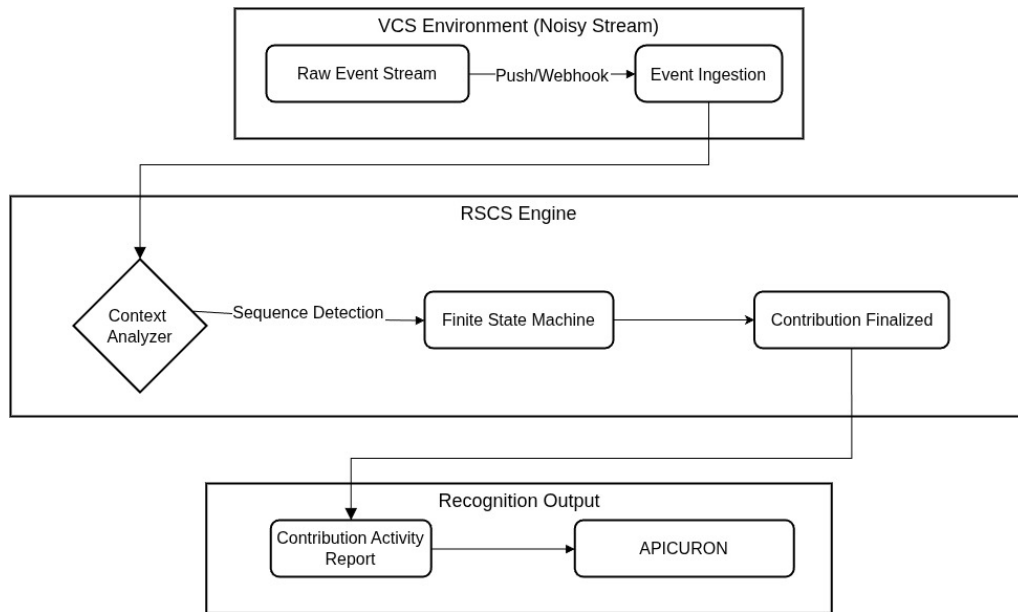


Figure 6.1: RSCS Data processing pipeline utilizing a Finite State Machine (FSM) to contextualize granular events into meaningful recognition outputs.

This approach to crediting software through an intermediary interpretation layer raises significant questions regarding governance and the interpretation of behavior. To address these, the RSCS is architected under three binding constraints:

- **Community Managed:** The definition of credit-worthiness is not universal. The goal is not to impose a rigid standard of what should be credited, but to provide a mechanism that allows specific research communities to define their own credit schema. The RSCS enables communities to configure patterns that match their values, ensuring the system remains relevant to domain-specific needs.
- **Versionable:** As community standards and development practices evolve, the definition of a contribution must adapt. The schema is designed to be versioned, allowing repository managers to update their recognition criteria over time without invalidating historical credit.
- **Transparency:** The transformation of a technical event into academic credit must never be a "black box." The model ensures that every attribution includes a provenance trace, providing a clear explanation of why a specific contributor was credited for a certain activity.

6.2.a Context-Aware Categorization

A fundamental principle of the RSCS is the rejection of the one-size-fits-all metric. Not all code is equal; a commit fixing a typo in documentation carries different semantic weight than a commit refactoring a core algorithm. Furthermore, the type of software being developed can be selected as a baseline to dictate the value of the contribution.

To resolve this, the RSCS proposes a recognition strategy that is aligned with the research software maturity model proposed by European Virtual Institute for Research Software Excellence (EVERSE). Leveraging the categorization of research software proposed and proposing a recognition strategy fitting for each category in the maturity model.

- **Analysis Code:**

This includes research software that captures computational research processes and methodology, and often occurs in the context of simulation, data generation, preparation, analysis and visualisation. This category of software is characterized by its almost exclusive usage by the author. Credit & recognition for this category focuses mainly on encouraging software discoverability and general software quality indicators with little focus on community contributions.

- **Prototype Tools:**

Refers to research software that demonstrates a new idea, method or model for use by others outside the project within which it originated, often as a substantive intellectual contribution in its own right and often in the form of a proof of concept.

Considering the broader use cases for this software, credit & recognition mechanisms in this category focuses on the collaborative aspects of software production, by recognizing the community support in the maturity model and how through categorized Pull Request submissions and active issue threads.

- **Research Software Infrastructure:**

This category encompasses libraries, databases, and platforms critical to the scientific community (e.g., DisProt, BioModels). Here, novelty is often less valuable than reliability. The crediting strategy in this category focuses on maintenance and uptime, heavily weighting contributions related to bug fixing, issue resolution, dependency management, and user support.

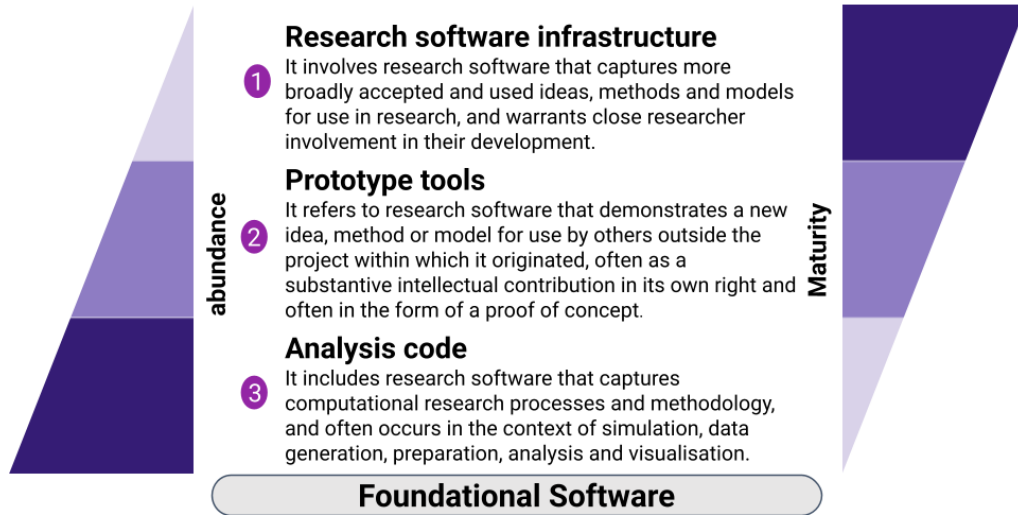


Figure 6.2: The EVERSE Software maturity model categorizing software into three categories depending on their maturity and abundance in the field.

6.3 Technical Implementation

The theoretical framework of the RSCS requires a robust technical baseline to function. A credit schema is only as valuable as the data it processes; therefore, the primary technical challenge lies in the accurate, reliable, and passive ingestion of research activity. This section details the Tracking Layer, the architectural component responsible for capturing software contributions without disrupting the researcher’s workflow.

6.3.a The Data Capture Challenge

The fundamental barrier to existing credit systems is the reliance on manual data entry. Systems that require researchers to manually claim or log their activities suffer from low adoption and rapid data staleness.

To mitigate this, the Tracking Layer is designed under a Zero-Friction constraint: the mechanism must capture events passively, as a side-effect of the researcher’s natural development workflow. The system must listen to the software repository’s stream of version control events requiring limited intervention from the researcher to trigger specific commands or modify their code to accommodate the credit system.

To achieve this passive monitoring within the GitHub ecosystem, three distinct integration mechanisms with GitHub have been evaluated: Webhooks, GitHub Actions, and GitHub Apps. Adopting the methodology for Real-time Event Monitoring, each approach is assessed against three critical criteria: Event Granularity, Security/Authentication, and Scalability.

- **Webhooks**

Webhooks provide a rudimentary “fire-and-forget” system where GitHub sends an HTTP POST payload to a configured server whenever an event occurs. While useful for simple notifications, they lack a robust authentication handshake and do not support the granular permission scopes required for reading historical data. Furthermore, managing webhooks across thousands of distinct repositories requires manual configuration for each, making this approach administratively unfeasible for large-scale infrastructure.

- **GitHub Actions**

GitHub Actions allow for the execution of workflows directly within the repository’s infrastructure. While effective for Continuous Integration (CI), using Actions for credit tracking inverts the control logic. It forces the repository maintainer to add specific YAML configuration files to their codebase. This violates the Zero-Friction constraint, as the credit system becomes a dependency that the researcher must maintain and debug.

- **GitHub Apps**

A GitHub App acts as a first-class actor within the ecosystem. Unlike a user or a script, an App can be installed across an organization or account with a single click. It possesses its own identity, enabling it to act independently of a specific user’s credentials.

Feature	GitHub Webhooks	GitHub Actions	GitHub Apps
Primary Use Case	External service notification (sending event payloads).	In-repository automation (e.g., CI/CD, scripting tasks).	Building robust, secure integrations across users/organizations.
Setup Cost	Medium. User configures in repo settings + manages external server.	Low. Users add a YAML workflow file; can use Marketplace to add actions.	Very Low. The user clicks “Install” from the Marketplace; the App handles the backend.
Development Cost	Low. Simple configuration and endpoint logic.	Medium. YAML workflow definition & runner environment knowledge.	High. Full application development (registration, key management, API logic).
Operation Cost	Low. Requires a small backend to receive events.	None - Low. Runs on GitHub servers (1k requests/hour/repo, 20 concurrent jobs, limited queues).	High. Runs on a dedicated infrastructure (web server).
Access Req.	Repository Admin/Owner for initial setup. None afterwards.	Repository Write Access for workflow files; temporary GITHUB_TOKEN.	Granular permissions granted by user/org during installation.
Data Scope	Limited. Only the event payload is provided.	Broad. Access to repository context, files, and API via GITHUB_TOKEN.	Broadest. Comprehensive API access based on granted permissions across installations.
Scalability	High. Scales with a robust backend, but high volume can strain the handlers.	Medium. Scales with GitHub’s runners (tiered limits) or self-hosted options.	Very High. Designed for platform-level integration across potentially thousands of installations.
Security	Payload validation via shared secret. Requires a secure endpoint.	Actions run in GitHub’s environment; uses scoped permissions and secrets.	Fine-grained permissions, short-lived tokens, and scoped repository access.

Table 6.1: Comparison of GitHub Integration Tools

6.3.b The Proof of concept

In the scope of this work, the github actions approach has been selected to produce the Commit Tracker. A proof-of-concept software built to establish a reliable method for formally recognizing and crediting software contributions. It operates in close connection with GitHub, taking advantage of features like Continuous Integration (CI) Continuous Deployment (CD), integration with ORCID, and the standardized development workflow provided by GitHub Actions. The system is designed to execute within a GitHub Actions environment, where it is triggered by GitHub and subsequently processes the repository that initiated the action.

The Commit Tracker serves as a proof of concept showcasing that data from GitHub Events can be submitted to APICURON. Therefore it employs a simple system that pushes all commits to APICURON as contribution events. Evidently, this ends up resulting in commit-counting but the purpose of the proof of concept here is to show that streams of events from platforms like github can be consumed, processed, and submitted to APICURON.

6.4 The Logic Layer: Stateful Contribution Mapping

While the Tracking Layer solves the problem of data ingestion, the raw event stream it produces is noisy and semantically shallow. A “Commit” event in isolation is ambiguous; it could be a typo fix, a failed experiment, or a critical algorithm refactor. Therefore, the core innovation of the RSCS lies in the Logic Layer, which transforms these discrete technical events into verified academic contributions.

6.4.a The Shift from Stateless to Stateful Recognition

Traditional software effort indicators often rely on non-representative metrics, such as raw counts of lines of code Lines Of Code (LOC) or number of commits. These are stateless indicators; they measure quantity, not impact. In the context of the RSCS, this approach is insufficient. Unlike Biocuration, where a single annotation often represents a complete atomic unit of work, software contributions are procedural and interdependent.

To address this, we introduce Stateful Contribution Mapping. This methodology treats a contribution not as a single event, but as a lifecycle that must traverse a specific sequence of states to be considered valid. The lifecycle in question is defined by the community and can be aligned with the proposed software

classification.

- The Atomic Unit of Work in this case is transient. A commit is merely a transitional state; it only gains academic value when it is peer-reviewed and integrated into the canonical codebase (the “Merged” state).
- Contextual Validity: Regardless of the measured target to identify a creditable unit of work. The validity of the measurement depends on the existence of the following characteristics indicating valuable human effort.
 - **Evidence** of intent and initiation of contribution: (e.g opening a Pull request)
 - **Iteration** (optional), New commits are added. Feedback is exchanged showing a continuous evidence of intent
 - **Validation**: approval of the contribution by automated tests or maintainers.
 - **Completion**: The contribution is integrated into the software.(e.g: commits are merged, PRs are merged. . . etc).

To implement this technically, the Logic Layer utilizes a Finite State Machine model where the github event stream doesn’t always trigger an event on the credit stream. But the logic layer keeps track of the state of the repository depending on the measured quality indicator.

In the example below we define a contribution to a feature as a fully solved closed PR that is tagged as “feature”. Using this definition we can configure the Finite state machine model to identify these patterns and react to them accordingly as shown in fig. 6.3.

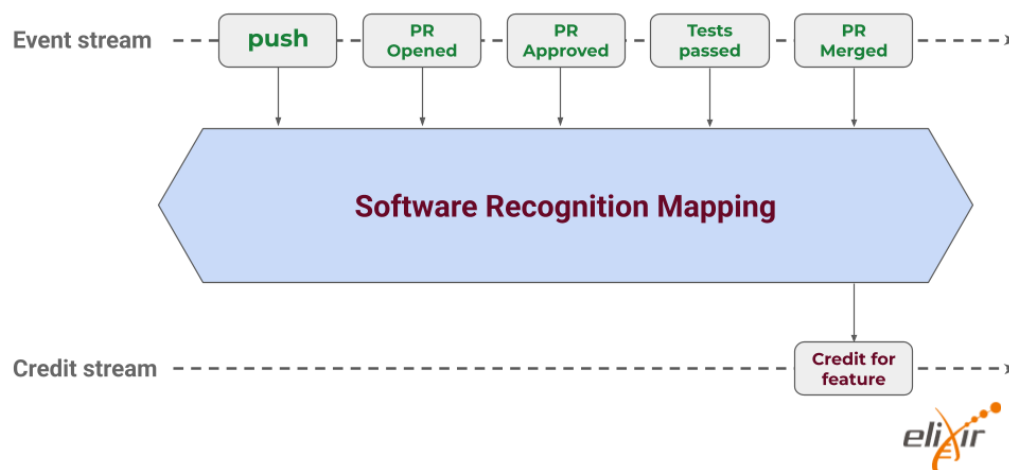


Figure 6.3: The Software Recognition Mapping captures all of the raw GitHub events but produces a credit stream only when the defined pattern is identified.

6.5 The Acknowledgement layer

The final component of the proposed architecture is the Acknowledgement Layer. Once a software contribution has been captured by the Tracking Layer and validated by the Logic Layer, it must be translated into a tangible academic asset. A valid credit entry stored in a private database has no value to a researcher's career; it must be propagated to the public infrastructure where research assessment occurs.

This section details the propagation mechanisms that feed verified software credits back into the global research ecosystem, ensuring interoperability with standard academic profiles and assessment frameworks.

6.5.a Ecosystem Integration and Feedback Loops

Following the architectural model defined in the Credit & Recognition Section, the RSCS functions as a specialized tracker that feeds into broader aggregator services. The aggregated Activity Reports are serialized into payloads and pushed to the APICURON API. This centralizes the data alongside other research activities (such as biocuration). Through APICURON's existing integration, these software contributions can be pushed to the researcher's ORCID profile, providing the researcher proper credit for their software contributions.

Part VII

Conclusions



Conclusions

This doctoral thesis set out to resolve a fundamental paradox in modern science: while research has become increasingly data-centric and software-reliant, the academic currency for excellence remains tethered to the static, narrative-driven publication. This misalignment has created a granularity gap, where the essential labour of biocuration and software engineering remains invisible and uncredited.

Through the design, development, and implementation of APICURON, this work has provided a robust technical solution to this sociotechnical problem.

Architectural Innovation: We established that static metadata is insufficient for capturing the dynamic nature of research. By implementing an event-driven architecture, APICURON successfully decoupled the execution of research tasks from their valuation, allowing for the capture of high-frequency, granular contribution events.

Validation in Biocuration: The implementation study across scientific resources such as DisProt and Reactome validated the APICURON model. We observed that while extrinsic crediting mechanisms (such as badges) had mixed results on curator motivation, the *visibility* provided by the platform proved essential for sustaining user engagement and stabilizing contribution workflows.

Sociotechnical Safeguards: The study revealed that technical tracking cannot exist without ethical safeguards. The development of the Consent Management System (chapter 5) addresses the *Consent Gap*, ensuring that granular tracking complies with privacy regulations GDPR without compromising the provenance record.

Extension to Software: Finally, we demonstrated that this event-driven model is domain-agnostic. By mapping the *Commit Fallacy* to the *Biocuration Gap*, we proposed the Research Software Credit Schema, laying the groundwork for treating software commits as first-class research outputs.

6.6 Implications for the Field

The findings of this thesis urge a shift in how the scientific community approaches infrastructure. We conclude that gamification is not a silver bullet for researcher engagement; rather, provenance and interoperability are the primary drivers of value.

For Biocuration, this means moving away from traditional authorship toward verifiable, acknowledged contributions that can be pushed to persistent profiles such as ORCID. For Research Software, it implies that we must move beyond citation files to tracking the process of engineering—maintenance, code review, and documentation—as distinct, creditable activities.

6.7 Perspectives

6.7.a From Gamification to Career Utility

While the implementation study revealed that professional curators are often indifferent to badges and medals, the value of the platform lies in its ability to synthesize granular actions into human-readable narratives. The perspective here is that credit systems must move away from gamifying work and toward recognizing expertise. Future iterations should focus on the automated generation of "Impact Portfolios" that researchers can directly utilize in hiring and promotion cycles.

6.7.b Semantic Middleware in a Federated Ecosystem

APICURON should not be perceived as a standalone repository but as a semantic middleware. The goal is to establish a federated ecosystem where disparate activity streams—from GitHub events to manual biocuration—are harmonized through a shared language of recognition. This involves strengthening the technical and legal bridges between contribution trackers and global PID providers like ORCID.

6.7.c Closing the Compliance and Trust Gap

The discovery of the Consent Gap highlights that technical interoperability is insufficient without regulatory interoperability. A future perspective for the field is the adoption of “Compliance-by-Design” infrastructures, such as the proposed Consent Management Platform. By making lawful data sharing the default rather than a hurdle, we can foster a more transparent and trustworthy research culture.

6.7.d The Human role in Data Production

As biology becomes increasingly reliant on Artificial Intelligence, the human investment in high-quality data becomes more critical. APICURON provides a mechanism to acknowledge the “human-in-the-loop” whose expert curation powers the next generation of AI breakthroughs. We foresee a perspective where attribution metadata becomes a quality indicator for the datasets used to train models, effectively closing the loop between the labor of curation and the utility of discovery.

Bibliography

- [1] “An Integrated Encyclopedia of DNA Elements in the Human Genome”. In: *PubMed* (2012). URL: <https://pubmed.ncbi.nlm.nih.gov/22955616/> (visited on 01/31/2026).
- [2] Michael Ashburner et al. “Gene Ontology: Tool for the Unification of Biology”. In: *Nature Genetics* 25.1 (May 2000), pp. 25–29. ISSN: 1546-1718. DOI: [10.1038/75556](https://doi.org/10.1038/75556). URL: https://www.nature.com/articles/ng0500_25 (visited on 01/20/2026).
- [3] Sucharitha Balu et al. “Complex Portal 2025: Predicted Human Complexes and Enhanced Visualisation Tools for the Comparison of Orthologous and Paralogous Complexes”. In: *Nucleic Acids Research* 53.D1 (Jan. 6, 2025), pp. D644–D650. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gkae1085](https://doi.org/10.1093/nar/gkae1085). URL: <https://academic.oup.com/nar/article/53/D1/D644/7903366> (visited on 01/31/2026).
- [4] Niall Beard et al. “TeSS: A Platform for Discovering Life-Science Training Opportunities”. In: *Bioinformatics* 36.10 (May 2020), pp. 3290–3291.
- [5] “Biocuration: Distilling Data into Knowledge”. In: *PLoS Biology* 16.4 (Apr. 16, 2018), e2002846. ISSN: 1544-9173. DOI: [10.1371/journal.pbio.2002846](https://doi.org/10.1371/journal.pbio.2002846). PMID: [29659566](https://pubmed.ncbi.nlm.nih.gov/29659566/). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5919672/> (visited on 12/16/2025).
- [6] claytonsiemens77. *Event Sourcing Pattern - Azure Architecture Center*. URL: <https://learn.microsoft.com/en-us/azure/architecture/patterns/event-sourcing> (visited on 03/27/2026).
- [7] CoARA – Coalition for Advancing Research Assessment. URL: <https://www.coara.org/> (visited on 01/30/2026).

-
- [8] *Data2health/Contributor-Role-Ontology: This Ontology Provides Contribution Roles for Use in Crediting Persons or Organizations*. GitHub. URL: <https://github.com/data2health/contributor-role-ontology> (visited on 01/15/2026).
- [9] Noemi del Toro et al. "The IntAct Database: Efficient Access to Fine-Grained Molecular Interaction Data". In: *Nucleic Acids Research* 50.D1 (Jan. 7, 2022), pp. D648–D653. ISSN: 0305-1048. DOI: [10.1093/nar/gkab1006](https://doi.org/10.1093/nar/gkab1006). URL: <https://doi.org/10.1093/nar/gkab1006> (visited on 01/31/2026).
- [10] European Organization For Nuclear Research and OpenAIRE. *Zenodo*. CERN, 2013. DOI: [10.25495/7GXX-RD71](https://doi.org/10.25495/7GXX-RD71). URL: <https://www.zenodo.org/>.
- [11] R. Fielding et al. *RFC2068: Hypertext Transfer Protocol – HTTP/1.1*. USA: RFC Editor, 1997.
- [12] David H. Freedman. *Wrong: Why Experts* Keep Failing Us—and How to Know When Not to Trust Them: *Scientists, Finance Wizards, Doctors, Relationship Gurus, Celebrity ... Consultants, Health Officials, and More*. New York, Boston, London: Little Brown & Co, 2010. 304 pp. ISBN: 978-0-316-02378-8.
- [13] Sara El-Gebali et al. "The Pfam Protein Families Database in 2019". In: *Nucleic Acids Research* 47.D1 (Jan. 8, 2019), pp. D427–D432. ISSN: 0305-1048. DOI: [10.1093/nar/gky995](https://doi.org/10.1093/nar/gky995). URL: <https://doi.org/10.1093/nar/gky995> (visited on 12/17/2025).
- [14] András Hatos et al. "APICURON: A Database to Credit and Acknowledge the Work of Biocurators". In: *Database* 2021 (Sept. 29, 2021), baab019. ISSN: 1758-0463. DOI: [10.1093/database/baab019](https://doi.org/10.1093/database/baab019). URL: <https://doi.org/10.1093/database/baab019> (visited on 12/17/2025).
- [15] Simon Hettrick. "The Hidden REF: Celebrating Everyone That Makes Research Possible". ModelShare Workshops. May 24, 2024. DOI: [10.5281/zenodo.11266576](https://doi.org/10.5281/zenodo.11266576). URL: <https://zenodo.org/records/11266576> (visited on 01/15/2026).
- [16] Diana Hicks et al. "Bibliometrics: The Leiden Manifesto for Research Metrics". In: *Nature* 520.7548 (Apr. 2015), pp. 429–431. ISSN: 1476-4687. DOI: [10.1038/520429a](https://doi.org/10.1038/520429a). URL: <https://www.nature.com/articles/520429a> (visited on 01/09/2026).
- [17] John Jumper et al. "Highly Accurate Protein Structure Prediction with AlphaFold". In: *Nature* 596.7873 (Aug. 2021), pp. 583–589. ISSN: 1476-4687. DOI: [10.1038/s41586-021-03819-2](https://doi.org/10.1038/s41586-021-03819-2). URL: <https://www.nature.com/articles/s41586-021-03819-2> (visited on 12/15/2025).

- [18] Rob Kitchin. “Big Data, New Epistemologies and Paradigm Shifts”. In: *Big Data & Society* 1.1 (Apr. 1, 2014), p. 2053951714528481. ISSN: 2053-9517. DOI: [10.1177/2053951714528481](https://doi.org/10.1177/2053951714528481). URL: <https://doi.org/10.1177/2053951714528481> (visited on 12/15/2025).
- [19] Rahuman S Malik-Sheriff et al. “BioModels — 15 Years of Sharing Computational Models in Life Science”. In: *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D407–D415. ISSN: 0305-1048. DOI: [10.1093/nar/gkz1055](https://doi.org/10.1093/nar/gkz1055). eprint: <https://academic.oup.com/nar/article-pdf/48/D1/D407/31698010/gkz1055.pdf>. URL: <https://doi.org/10.1093/nar/gkz1055>.
- [20] Bálint Mészáros et al. “Minimum Information Guidelines for Experiments Structurally Characterizing Intrinsically Disordered Protein Regions”. In: *Nature Methods* 20.9 (Sept. 2023), pp. 1291–1303. ISSN: 1548-7105. DOI: [10.1038/s41592-023-01915-x](https://doi.org/10.1038/s41592-023-01915-x). URL: <https://www.nature.com/articles/s41592-023-01915-x> (visited on 01/24/2026).
- [21] Bertrand Meyer. *Object-Oriented Software Construction*. New York: Prentice Hall Press, 1994. 592 pp. ISBN: 978-0-13-629049-0.
- [22] Marija Milacic et al. “The Reactome Pathway Knowledgebase 2024”. In: *Nucleic Acids Research* 52.D1 (Jan. 5, 2024), pp. D672–D678. ISSN: 0305-1048. DOI: [10.1093/nar/gkad1025](https://doi.org/10.1093/nar/gkad1025). URL: <https://doi.org/10.1093/nar/gkad1025> (visited on 01/31/2026).
- [23] David Moher et al. “The Hong Kong Principles for Assessing Researchers: Fostering Research Integrity”. In: (). DOI: [10.1371/journal.pbio.3000737](https://doi.org/10.1371/journal.pbio.3000737). URL: <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000737> (visited on 01/09/2026).
- [24] Sam Newman. *Monolith to Microservices: Evolutionary Patterns to Transform Your Monolith*. Beijing China ; Sebastopol, CA: O’Reilly Media, 2019. 270 pp. ISBN: 978-1-4920-4784-1.
- [25] NISO CRediT Working Group. *ANSI/NISO Z39.104-2022, CRediT, Contributor Roles Taxonomy*. DOI: [10.3789/ansi.niso.z39.104-2022](https://doi.org/10.3789/ansi.niso.z39.104-2022). URL: <http://www.niso.org/publications/z39104-2022-credit> (visited on 01/30/2026).
- [26] Node.js — *Run JavaScript Everywhere*. URL: <https://nodejs.org/en> (visited on 01/31/2026).
- [27] Maria Victoria Nugnes et al. “DisProt in 2026: Enhancing Intrinsically Disordered Proteins Accessibility, Deposition, and Annotation”. In: (). URL: <https://dx.doi.org/10.1093/nar/gkaf1175> (visited on 01/19/2026).

- [28] *Open Science - Research and Innovation - European Commission*. Dec. 17, 2025. URL: https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science_en (visited on 01/10/2026).
- [29] Sylvain Poux et al. "On Expert Curation and Scalability: UniProtKB/Swiss-Prot as a Case Study". In: *Bioinformatics* 33.21 (Nov. 1, 2017), pp. 3454–3460. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btx439](https://doi.org/10.1093/bioinformatics/btx439). PMID: [29036270](https://pubmed.ncbi.nlm.nih.gov/29036270/). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5860168/> (visited on 12/17/2025).
- [30] *PROV-O: The PROV Ontology*. URL: <https://www.w3.org/TR/prov-o/> (visited on 12/18/2025).
- [31] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2025. URL: <https://www.R-project.org/>.
- [32] *Read the Declaration*. DORA. URL: <https://sfdora.org/read/> (visited on 01/30/2026).
- [33] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance)*. May 2016.
- [34] *Restify*. Restify. URL: <https://restify.github.io/> (visited on 01/20/2026).
- [35] *SCoRO, the Scholarly Contributions and Roles Ontology*. URL: <https://sparontologies.github.io/scoro/current/scoro.html> (visited on 01/15/2026).
- [36] D. Sculley et al. "Hidden Technical Debt in Machine Learning Systems". In: *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html (visited on 12/16/2025).
- [37] Arfon M. Smith et al. "Software Citation Principles". In: *PeerJ Computer Science* 2 (Sept. 19, 2016), e86. ISSN: 2376-5992. DOI: [10.7717/peerj-cs.86](https://doi.org/10.7717/peerj-cs.86). URL: <https://peerj.com/articles/cs-86> (visited on 01/15/2026).
- [38] Anne E. Thessen et al. "Proper Attribution for Curation and Maintenance of Research Collections: Metadata Recommendations of the RDA/TDWG Working Group". In: *Data Science Journal* 18.1 (Nov. 8, 2019). ISSN: 1683-1470. DOI: [10.5334/dsj-2019-054](https://doi.org/10.5334/dsj-2019-054). URL: <https://datascience.codata.org/articles/dsj-2019-054> (visited on 12/18/2025).

-
- [39] Gustavo F. Tondello et al. “The Gamification User Types Hexad Scale”. In: *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY ’16. New York, NY, USA: Association for Computing Machinery, Oct. 15, 2016, pp. 229–243. ISBN: 978-1-4503-4456-2. DOI: [10.1145/2967934.2968082](https://doi.org/10.1145/2967934.2968082). URL: <https://dl.acm.org/doi/10.1145/2967934.2968082> (visited on 01/24/2026).
- [40] Fernando van der Vlist, Anne Helmond, and Fabian Ferrari. “Big AI: Cloud Infrastructure Dependence and the Industrialisation of Artificial Intelligence”. In: *Big Data & Society* 11.1 (Mar. 1, 2024), p. 20539517241232630. ISSN: 2053-9517. DOI: [10.1177/20539517241232630](https://doi.org/10.1177/20539517241232630). URL: <https://doi.org/10.1177/20539517241232630> (visited on 12/16/2025).



Acknowledgments

I wish to begin by expressing my deepest gratitude to my parents, whose unwavering belief in my potential has been the cornerstone of this doctoral journey. Their countless sacrifices ensured that I was provided with the optimal conditions for my academic and personal growth, and for this, I am eternally grateful. To my family and friends, your continued faith in my capabilities has been an invaluable source of strength; I am forever thankful for your presence in my life.

I would like to extend my sincere appreciation to my esteemed supervisors, Prof. Silvio Tosatto and Dr. Federica Quaglia. Their relentless guidance, technical expertise, and encouragement have been instrumental at every stage of this research. I am profoundly grateful for their dedication to bringing out the best in my work. I also thank my co-supervisor, Dr. Luciano Gamberini, for his support.

I am indebted to my colleagues at the University of Padua, with whom I have had the privilege of collaborating over the last three years. The shared experiences and memorable moments within the department will be cherished forever. My thanks also go to the members of the jury for their time, patience, and insightful examination of this thesis.

A significant portion of this work's maturity is attributed to the early adopters who participated in the implementation study. I am particularly grateful to the teams at EMBL-EBI—specifically those representing Pfam, IntAct, Complex Portal, Reactome, and BioModels for their willingness to integrate APICURON and for providing the critical feedback that drove the system's technical and regulatory evolution.

This research was supported by the financial contribution of the European Union NextGenerationEU through the National Recovery and Resilience Plan (PNRR), M4C2 Investment 3.1. I also acknowledge the "Elixir x NextGenerationIT" program for its role in strengthening the Italian research infrastructure.

Finally, I offer my heartfelt appreciation to every individual who offered support in any form, however small. Your gestures made a definitive difference in the completion of this work.