Decentralized Research Data Management: Introducing SoVisu+

Abstract:

Introduction:

Effective research data management serves as the cornerstone for scientific progress, enabling informed decision-making and driving innovation. However, centralized and siloed approaches, prevalent in many current research information systems (CRIS), often face challenges in ensuring data quality, accessibility, and community engagement. SoVisu+, building upon the foundations of SoVisu (Reymond 2022), proposes a novel decentralized data management model that addresses these limitations, paving the way for a more efficient and sustainable research environment.

Challenges of Centralized Systems:

France's research data landscape exemplifies the inherent shortcomings of centralized systems. Research data, encompassing diverse aspects like laboratory descriptions, financial management, human resources, and production management, is often complex, dispersed across disparate databases, and unevenly distributed across institutions. This fragmented data structure lacks effective interoperability, leading to cumbersome and repetitive efforts for information retrieval. Researchers, forced to utilize expensive external bibliographic databases and reconcile fragmented data sets, often resort to manual identification and confirmation of research outputs, fostering a sense of resignation and hindering trust in data accuracy. Additionally, relying on static data silos leads to information obsolescence, as research work is inherently dynamic, and its descriptions evolve over time.

Empowering Researchers and Institutions:

SoVisu+ fundamentally shifts the paradigm by empowering researchers through self-archiving tools and real-time visualizations of their research profiles (Reymond et Tabariès 2021 ; Reymond, Tabariès et Bara 2022). By showing each researcher a mirror of the lexical representation of the production associated with their identifiers (Figure 1), the researcher is empowered regarding their discoverability in indices. Each can then take ownership and decide to improve the readability of their scientific profile. This fosters a sense of ownership and encourages researchers to actively curate and improve the quality of their data. Institutions, capitalizing on this expert-verified information, can generate reliable and relevant indices for various purposes such as management, evaluation, and valorization. This not only streamlines administrative processes but also fosters a culture of transparency and accountability within the research ecosystem. SoVisu+ thus opens a path for instrumented dialogue with researchers to collectively curate data, obtain qualitative information, and non-public information. In exchange for the provided tools and under the auspices of certification authorities, the community contributes to reconstructing and verifying the knowledge graph of affiliation entities and recursively up to the institution level.

Building a Collaborative Knowledge Graph:

SoVisu+ fosters collaboration by enabling the community to collectively curate data, contribute to the construction of a robust knowledge graph (KG), and provide internal information. This KG leverages Semantic Web and Social Web standards alongside AI/LLMs to facilitate data management tasks, eliminate redundancy, and surpass the capabilities of existing CRIS. Researchers and authorized entities work collaboratively to populate and enrich the KG, ensuring its accuracy and coherence with the evolving scientific activity. Information flows are managed at the institutional level within the SO ecosystem to ensure deposit quality, community adherence, and data integrity.

Expanding Functionality and Services:

SoVisu+ adopts a continuous agile development approach, gradually expanding its functional scope to encompass diverse research products such as dissertation thesis, patents, and preprints. This comprehensive approach ensures that the system caters to the diverse needs of researchers across various disciplines and research stages. Additionally, it progressively reconstructs CVs and offers AI-powered tools to facilitate metadata application and model improvement through usage data. By offering a comprehensive suite of services (Reymond et Lapôtre 2023), SoVisu+ empowers researchers to save time, manage their data efficiently, and ultimately enhance the dissemination and discoverability of their research outputs. Through incremental functional scope expansion, an agile and participative development mode, we will gradually extend to general research activities to establish a participative institutional management device (Boukacem 2023).

Benefits for Researchers and the Community:

By utilizing SoVisu+'s services, researchers will benefit from several key advantages. Time saved through streamlined data management processes allows them to focus on core research activities. Additionally, the system ensures data security and validation, fostering trust and confidence in the research information ecosystem. Furthermore, improved research dissemination and discoverability lead to increased research visibility and potential research collaborations. The community, through collective effort, reconstructs and maintains the "knowledge capital" of affiliated entities, contributing significantly to institutional knowledge graph creation. Transparently leveraging Semantic Web technologies, the community naturally rebuilds the 'institutional knowledge graph' and certifies its quality to strengthen it, consolidating the topology of data certification and validation points once and for all: this graph can then feed public portals, websites, or institution expertise search tools, etc. with measurable accuracy and, by extension, evaluation reports, production monitoring, and management at each node of the hierarchical organisation…if any.

Formats and Interoperability

The researcher interface of SoVisu+ serves as a "consultation" or indirect mediation place with experts: the basic predicate is that they know their affiliations, productions, expertise, discipline(s), etc., and can attest to or preferably validate this information simply, then perhaps highlight certain aspects. In this approach, researchers are inherently authorities on their scientific profiles. These data are then deposited in a container that can present different views (what is private, public, relevant for institutional certification, etc.), information appreciated by the expert, possibly supplemented and, if necessary, explained. The description of these contents, as much as possible, will follow one of the international standards (CERIF or VIVO (Krafft *et al.* 2010 ; Ilik *et al.* 2018 ; Jörg, Höllrigl et Baker 2014 ; Kremenjaš,

Udovicic et Orel 2020) with a guarantee of continuity because these models are extensible and structured to cover research activity in the broad sense (cf. Figure 1) in an evolving format. The use of one of these formats and specialized ontologies for descriptive metadata is also a necessary condition to ensure interoperability not only at the national but also international level. Highlighting this profile on institutional portals, expert identification systems, bibliometric reports, or even evaluation devices ensures that they are likely to be carefully reviewed to the expert's standard, who remains the master of what is disseminated. The system should also gradually support research activities: a new research project from its definition, submission, evolution will be under the authority of the profile. Some parts will fall under certification authority regarding data validation following the same principle as before. Its description, compatible with the institutional repository, will thus not only allow monitoring but also ensure reliability by adhering to its evolutionary dynamics.

Figure 1, below, shows the scope of the CERIF model (Jörg, Höllrigl, and Baker 2014), continuously maintained to evolve with research management needs. At the heart of this model, Organizational Units (Laboratories, Groups, ...) are represented just like people or projects, and if necessary, profiles, productions, infrastructures, activities, and results, as well as indicators, are attached. This formalized representation in interoperable and machine-readable standards relies on long-lasting identifiers as much as possible for tracking.

Decentralized Web Technologies:

SoVisu+ embraces recent advancements in decentralized web technologies to empower individual researchers by leveraging the principles of data ownership and control. The system utilizes Solid Pods, Activity Streams, and Linked Data (LDN) to facilitate real-time data capture, automated notifications for certification workflows, and knowledge graph construction. This innovative approach empowers researchers to manage their data securely and fosters a sense of agency within the research ecosystem.

Knowledge Graph Reconstruction Strategy:

The knowledge graph is progressively built through a well-defined approach involving initial data population from existing systems followed by researcher-driven updates and certifications. This strategy leverages a combination of automated and collaborative elements to ensure data accuracy and completeness. We do not provide an exhaustive overview here but emphasize recent works by Tim Berners-Lee on the Solid technology ("Solid-Based Approach for Research Information Interoperability - General Discussion" 2023). Scientific communication needs to be rethought (Sompel et al. 2004; Sponberg et al. 2023), and works in this domain remain highly active. Among other projects targeting decentralization in scientific production (e.g., OnePub), the Researcher Pod project, funded by the Mellon Foundation, explores a collaboration system based on social web protocols to instrument various stages. The system is based on Tim Berners-Lee's decentralized web vision and the various related Solid technology protocols (security, authenticated, perennially identified, ...) as envisioned by Capadisli (2020) and Van de Sompel (Sompel et al. 2004) for users. In the context of the Solid project, a Solid Container aims to give individuals control over their own web data. It provides a personal online storage space where users can store and manage their

data, then grant permissions to applications or other users to access it. Linked Data Platform servers, in this context, are used to implement the server-side of these Solid Containers, enabling users to manage and share their data according to the principles of Solid and Linked Data. Solid Containers can thus be used in a wide range of applications (Verstraete, Verbrugge, and Colle 2022) where individuals want to maintain ownership and control over their personal data (Pandit 2023; Solanki 2021). This includes social networks, data sharing, collaborative workspaces, and much more. Transposed to the research domain, the work of Langer et al. describes specifications for a "research project" type container (Langer, Vu Nguyen Hai, and Gaedke 2020).

A network of Researcher Containers (Researcher Pod) is decentralized: each researcher references research results in their personal data container (Mansour et al. 2016; Sambra et al. 2016) hosted on a personal web domain. This action, and its updates, alert automated services or network actors of these modifications. Value is thus added without further action from the initial user to these research results through interaction with appropriate decoupled services (actors or robots). Specifications defined in the Activity Stream protocols ("Activity Streams 2.0" 2017) are an excellent example of this capacity. In the context of scientific communication, in a decentralized web, these containers guarantee the sustainability and independence of the data and contribute to the development of a decentralized ecosystem for research results. A new generation of decentralized, secure, and scalable systems and protocols is thus emerging, responding to the need for a fair and sustainable scholarly communication system. In this context, we extend this principle currently specified for scientific publishing to other research activities, from basic affiliation elements to other details of scientific productions and their ecosystem: projects, theses, events. A set of elements that, ultimately, will cover scientific activity, converging the interface into both a CV management and an interface for exchange and mediation with the ecosystem (documentation, HR, finances, ...). Business silos will be able to point to this data to avoid duplication, redundancy, and data inconsistency. Beyond these technical details, which will be subject to open specifications meeting W3C requirements, we focus here on the integration strategy within an institution that conditions the coherence of the system via researcher identifiers, particularly the alignment of the quadruplet (WebId, idhal, OrcId, Idref) and affiliation with a research entity (group, laboratory, ...) and their PID (De Castro *et al.* 2022 ; de Castro *et al.* 2023 ; De Castro *et al.* 2023)after passing through a certification authority. The combination of the three technologies Pod Solid, Activity Stream, and Linked Data (LDN) pave the way for the capture of research data (productions, activities, etc.) opened by the researcher to rights holders and in real time: the Containers inform robots of any changes (additions, deletions, and updates), which refer to authority nodes (laboratory manager, documentation services...), for certification or validation without any further action necessary. The Knowledge Graph realized through recursive aggregation is collectively maintained in coherence with the progress of scientific activity.

Researchers Interface:

SoVisu+ prioritizes user experience by providing a user-friendly interface dedicated to individual researchers. This interface offers valuable services, simplifying data management and facilitating data reuse across various contexts. Inspired by an agile design philosophy, the interface integrates functionalities like CV management, scientific production curation, and support for metadata generation (automatic extraction from text, standardized vocabularies, and AI assistance). Additionally, the system facilitates the archiving of research outputs with a single click. Looking towards the future, SoVisu+ aims to incorporate functionalities like

project creation linked with research services, event creation, application of signature charters, and seamless integration with laboratory and colleague data, ultimately creating a one-stop shop for researchers to manage all aspects of their research activities. These functionalities, coupled with automated notification and certification processes, add significant value to the research ecosystem. For researchers, this translates to streamlined data management, for the ecosystem, it fosters informed decision-making, and the entire system seamlessly points to the researcher's Container, where they remain in control of their data. This centralized location serves as a repository and exchange hub, paving the way for a more transparent and open approach to research evaluation and steering.

Rethinking Evaluation and Steering:

Bibliometrics has traditionally occupied a central role in research evaluation and steering. While vital in understanding research dynamics and identifying emerging trends, it faces limitations. Its reliance solely on publications paints an incomplete picture of academic contributions, and the coverage and consistency of bibliographic databases raise concerns about data accuracy and fairness. Furthermore, comparisons across diverse disciplines can be misleading, and focusing solely on quantitative indicators risks overlooking qualitative aspects of research.

SoVisu+ aims to address these limitations by introducing a more comprehensive and transparent approach to evaluation and steering. By linking self-archiving with CV management and evaluation, researchers are empowered to provide valuable insights into their diverse activities beyond publications matching DORA compliance (Schöpfel et Azeroual 2022) and most Leiden principles. This includes considerations for:

- **Research not captured in traditional metrics:** Activities like data collection, public outreach through media creation, and teacher-researcher contributions that enrich society beyond traditional publication channels.
- **Historical context:** The ability to track the evolution of research production within an organization over time provides valuable insights into strategic directions, scientific advancements, and the cumulative impact of the organization's research efforts.
- **Represent ethical practice:** at least partially (Schöpfel et Azeroual 2023)
- **Transparency and mediation:** Researchers can review and provide context surrounding quantitative indicators, allowing for a more nuanced understanding of their contributions. Additionally, the system encourages open dialogue between researchers and evaluation bodies, facilitating a more collaborative approach to evaluation as a customizable solution proper to each institution.

Conclusion

By leveraging the power of decentralized technologies and fostering community collaboration, SoVisu+ empowers researchers, improves data quality and accessibility, and fosters a more transparent and inclusive research ecosystem. This innovative approach paves the way for streamlined data management, improved research evaluation practices, and ultimately, accelerates scientific progress through informed decision-making and collaborative knowledge creation. Aside the cost of such architecture, the implementation will need to be carefully adapted to the ecosystem: years of practices in centralized silos that were built for security reasons, business and technologies choices have conditioned historical behaviour and expectations of the knowledge graph. In the proposed architecture missing

users (and their production) from the community are mechanically out of the system, excluded from disseminations views, assessments processes and so on. In a certain way this comes from an individual choice that can be understood and, reciprocally this tends to a better consistence of the resulting graph and monitoring indicators.

Boukacem, C. 2023. « La réforme de l'évaluation de la recherche ». Dans . https://mediaserveur.u-bourgogne.fr/videos/5-la-reforme-de-levaluation-de-la-recherche/.

Castro, Pablo de, Ulrich Herb, Laura Rothfritz et Joachim Schöpfel. 2023. « The role of universities in the implementation of persistent identifiers (PIDs) ». Dans *Proceedings of european university information systems congress 2023*, 95 : 291-300. Vigo, Galicia, Spain. https://doi.org/10.29007/97w6.

De Castro, Pablo, Ulrich Herb, Laura Rothfritz et Joachim Schöpfel. 2022. « Some Reflections on the Current PID Landscape – with an Emphasis on Risks and Trust Issues ». *Procedia Computer Science* 211 : 28-35. https://doi.org/10.1016/j.procs.2022.10.173.

—. 2023. « The gradual implementation of organisational identifiers (OrgIDs) ». Zenodo. https://doi.org/10.5281/ZENODO.7327535.

Ilik, Violeta, Michael Conlon, Graham Triggs, Marijane White, Muhammad Javed, Matthew Brush, Karen Gutzman, *et al.* 2018. « OpenVIVO: transparency in scholarship ». *Frontiers in Research Metrics and Analytics* 2. Frontiers : 12.

Jörg, Brigitte, Thorsten Höllrigl et David Baker. 2014. « Harmonising and Formalising Research Administration Profiles CASRAI / CERIF ». *Procedia Computer Science* 33 : 95-102. https://doi.org/10.1016/j.procs.2014.06.016.

Krafft, Dean B, Nicholas A Cappadona, Brian Caruso, Jon Corson-Rikert, Medha Devare, Brian J Lowe, VIVO Collaboration, et others. 2010. « Vivo: Enabling national networking of scientists ». Dans *Web science conference*. Raleigh, NC, USA.

Kremenjaš, Davorin, Petra Udovicic et Ognjen Orel. 2020. « Adapting CERIF for a national CRIS: A case study ». *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, 1633-1638.

Reymond, David. 2022. « Involving researchers in Open Science: the SoVisu innovative solution ». Dans *EOSC Symposium, session MVE beyond EOSC Future and the INFRAEOSC07*. Prague, Tchéquie.

Reymond, David et Raphaëlle Lapôtre. 2023. « SoVisu+: starting point and foundations of a national CRIS ». Dans *Semantic Web in Libraries*. Berlin / Germany.

Reymond, David et Alaric Tabariès. 2021. « Le dispositif SOVisu ». Python. Université de Toulon.

Reymond, David, Alaric Tabariès et Lena Bara. 2022. *Accompagnement et visualisations de la science ouverte (SoVisu)*. https://hal.science/hal-04103412.

Schöpfel, Joachim et Otmane Azeroual. 2022. « What does DORA mean for CRIS? » Dans *euroCRIS strategic membership meeting 2022*. Nijmegen, Netherlands : euroCRIS. https://hal.univ-lille.fr/hal-03883806.

—. 2023. « Ethical issues of the organization and management of research information ». Dans *Fourth international conference on the ethics of information & knowledge organization*. Lille, France : ISKO International Thematic Network on Ethics in SHS. https://hal.univ-lille.fr/hal-04291025.