Delivery of a proof of concept for terms4FAIRskills

Report on #Terms4FAIRskills project (no. 074) to European Open Science Cloud (EOSC) Co-creation Fund, Sep 2020 - Feb 2021

March 2021

Authors: Yann Le Franc, Peter McQuilton, Laura Molloy

This activity has been produced with the support of EOSCSecretariat.eu, which has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-2018-4, Grant Agreement number 831644.

Table of contents

1. Introduction and aims

Science is at a crossroads. More and more data is being produced, but we do not yet have the tools, services and technical capacity to use this data to its full potential. One major stumbling block is the provision of educational and training materials to those creating and maintaining data today. While there are some training materials and courses now coming online, how does a user find them, or know what they will learn from them? This is where terms4FAIRskills comes in.

Through this EOSC Co-creation award, we aim to deliver a proof of concept terminology to capture the skills and competencies necessary to make and keep data FAIR. This terminology will enable the cross domain and cross-repository searching for training materials by the skills and competencies they require and confer.

We foresee, once mature, that this terminology will apply to a variety of use cases, including but not limited to the:

- creation and assessment of stewardship curricula;
- annotation, discovery and evaluation of FAIR-enabling materials (e.g. training) and resources;
- formalisation of job descriptions and CVs with recognised, structured competencies.

The completed terminology will be of use to trainers teaching FAIR data skills, researchers who wish to identify skill gaps in their teams, and managers who need to recruit individuals with specific skills and competencies, and to develop FAIR competency-based reward structures and training plans.

On a strategic level, these aims are confluent with the overall ambitions of the EOSC. *Digital Skills for FAIR and Open Science* (2021) identifies ten actors (including their roles and related skills) in the EOSC ecosystem for whom skills and training is relevant (see Figure 1). The mature terminology will be an asset for these audiences in the use cases listed above, as EOSC implements its vision of a web of FAIR data and related services for science, making research data interoperable and machine actionable following the FAIR guiding principles[1].

[1]https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/european-open-science-cloud-eosc_en

Figure 1: Ten EOSC actors. From [Digital skills for FAIR and Open Science : report from the EOSC Executive Board Skills and Training Working Group.](), p. 17. Used under licence CC-BY.

Terms4FAIRskills is potentially of use to all ten actors in their engagement with FAIR data. We think it will be primarily of use to:

- Educators/Trainers in the annotation of their training materials and for discovery of their training materials.
- Researchers in their skills development and, when mentoring or managing other researchers, in the creation and assessment of training plans.
- Data Curators, Data Stewards/Data Librarians, Data Scientists/Data Analysts in the annotation of their skills sets, the presentation of these skills in CVs or otherwise, and the assessment of their Continuing Professional Development (CPD) activities.

2. Activities

Terms4FAIRskills was born from a community need, identified at the GO FAIR Coordination meeting in January 2019. We hosted two further workshops where the scope and purpose of a terminology for FAIR skills was developed, and we began laying plans for ontology creation. Further detail of each workshop is available here:

- The first f2f meeting (Paris, FR May 2019)
- The second f2f meeting (The Hague, NL October 2019)

It was clear that there was great potential for a focused effort to bring this idea to a pilot, minimum viable product to demonstrate the utility of such a terminology. The EOSC Co-creation award allowed us to focus on this process.

The EOSC Co-creation award ran from 22 Sep 2020 to 22 Feb 2021 and funded a core team (see Appendix 4) who produced a programme of activities that exceeded the initial ambitions of the project workplan. The workplan aimed to hold one core team sprint and three stakeholder events with participation from the wider community. In fact, we conducted a more extensive series of activities, including participatory events and outreach events, as listed below.

2.1. Participatory hack series

- Three core team hacks - where we developed the terminology, iterated and improved the conceptual model, assessed the use of the Semaphora annotation tool (a Google Chrome plugin derived from B2NOTE[2]) and incorporated a number of terms from relevant existing ontologies.
- Five group hack sessions - 11 and 15 Dec 2020; 27 and 29 Jan 2020, and an 'annotate-a-thon' hack week (comprising a demo, discussion and review session on 22 Feb plus two further drop-in sessions on 24 and 26 Feb 2021). These hack sessions were run in partnership with a wide range of interested stakeholders in the RDM and FAIR data sectors including participants from the ELIXIR training platform, the FAIRsFAIR project, the CODATA-RDA schools of Research Data Science, the EOSC-Pillar project, the Digital Curation Centre, DANS, and CINES (as well as representation in the core team from CODATA and FAIRsharing). Two members of the participant group are also members of the EOSC Skills & Training Working Group. In these sessions, we focused on applying the terminology to materials volunteered by the participants and in so doing, introduced people to the project, took them through our conceptual thinking as encapsulated by the model, discussed the use cases, introduced the notion of applying the terminology to material via the concept of annotation and demonstrated the annotation process using the Semaphora Chrome plugin. The final 'annotate-a-thon' hack week served as an extended workshop to deliver a demo, final review and discussion of next steps to wrap up this phase of activity.

2.2. Outreach and communications activities

We engaged in a large number of outreach, promotion, and communications activities during the funded period. Below, we list the main external events. (Links to outputs from all events are listed in section 4, 'Publicly available outputs'.)

- 22nd October 2020: EOSC Symposium (presentation);
- 30 November and 2 December 2020 (2 iterations): GO-FAIR/CODATA FAIR Convergence Symposium, panel on 'FAIR Data Stewardship: training and career opportunities' (presentation and panel discussion);
- 12th February 2021: ELIXIR Converge meeting (presentation);
- 22nd February 2021: ELIXIR EDAM and FAIR terminology (presentation);

[2] https://b2note.eudat.eu

- 28th February 2021: ELIXIR FAIR Training Working Group (presentation);
- Twitter engagement with hashtag #terms4FAIRskills;
- Blogposts and news articles on the CODATA blog, FAIRsharing project blog, FAIRsFAIR project news page.

3. Outputs and findings

The pilot project created a range of outputs and findings.

3.1. Deliverables

As specified in our proposal, our two key deliverables were:

- D1: Develop the terminology with the community to V0.1;
- D2: Create a proof-of-concept prototype for cross-use case annotation, categorisation, searching and display of FAIR skills courses.

Development of the terminology and proof-of-concept prototype was driven by working with training materials in current use. Our intended material providers were ELIXIR, through the ELIXIR Training and Events Portal (TeSS), the CODATA-RDA Schools of Research Data Science in partnership with FAIRsFAIR, and the GO FAIR curriculum for FAIR data, taught at the University of Leiden. However, after extended consultation with the GO FAIR team, it was established that they were unable to share their materials at this point due to licensing complexities. We adapted to work with an extended group of stakeholders including the ELIXIR-FR, ELIXIR-NL and CODATA-RDA School of Research Data Science teams and participants from DANS, CINES and the FAIRsFAIR project (specifically WP6 which is focused on the creation of a competence centre for FAIR data).

The terminology underwent considerable revision in response to the feedback from our stakeholder groups and from the questions and comments yielded by the series of hack sessions. In the first phase of the development (commencing Sept 2020), we started evaluating the 243 concepts that were developed during the various workshops prior to the EOSC Secretariat Award. These had been captured primarily using the online WebProtege ontology management tool.

As many concepts were ambiguous and/or without proper definitions, we underwent a long process of refining the core classes of the model and the logic in order to answer a set of competency questions that were defined to achieve the aims of this proposal. This allowed us to redefine the terminology model (referred to as core model v1 in our presentations) and remove a large number of terms that were not fit for purpose. The initial model represents Knowledge, Skills and Aptitudes (KSAs) as core classes with a large number of more specific subclasses.

After this first phase, we investigated existing additional sources of concepts for describing the Knowledge, Skills and Aptitudes appropriate for data stewards (including the EDISON and FAIR4S frameworks and the CASRAI RDM Glossary) in order to evaluate potential alignment and reuse of existing ontological sources. In parallel, we tested the updated version of terms4FAIRSkills by annotating one of the training materials sets made available by our partners.

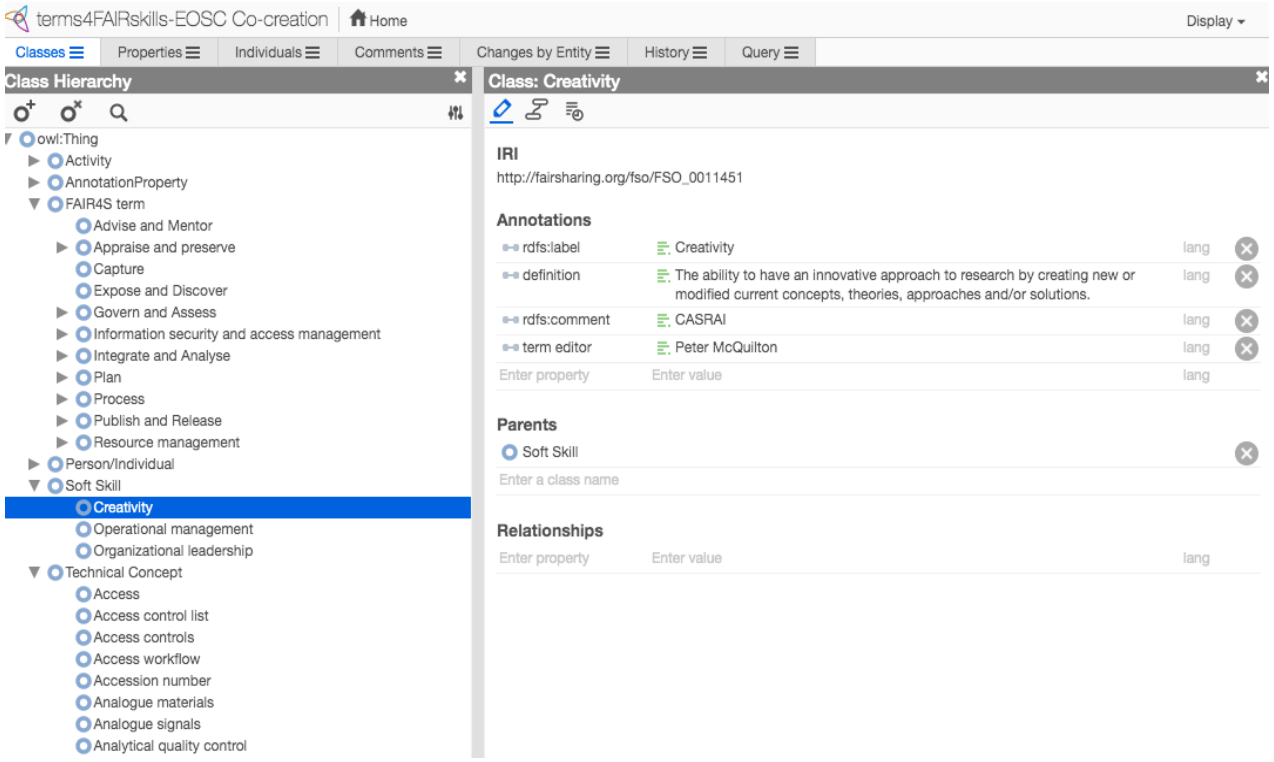

Figure 2: Working on the terminology in Webprotege. Each term has an IRI (International Resource Identifier) and where reliable definitions are available, these are noted.

Through this process, we identified a number of issues associated with the annotation process. In particular, we identified the need to separate technical concepts from other classes and to completely redesign the underlying model of the ontology. The main reason for this change is the difficulty to directly annotate training materials containing technical information with either knowledge and skills about specific technical aspects. In doing so, this would have created significant redundancy and duplication in the terminology. This test concluded the second phase of the development.

In the third phase, we updated the model to integrate 346 technical concepts from the CASRAI RDM glossary and refined the logical model to represent Knowledge, Skills and Aptitudes as being derived from the training material describing technical concepts. This resulted in the terminology having 746 terms. This change is reflected in the schematic representation of the new version of the model shown in figure 3.

Once these major changes have been implemented, we started the fourth phase of development, testing the ontology with domain experts to capture new terms and benchmark the current ontology with concrete use cases. We used an agile-like iterative approach that leverages the Semaphora Chrome plugin derived from the semantic annotator B2NOTE, a EUDAT/EOSC-Hub service. The plugin allows the user to annotate content on the Web with ontologies and provides a means to collect missing concepts in the ontology from domain experts, using annotations with free text keywords. It has been used to annotate Google presentations, Zenodo records and TeSS records. Semaphora creates annotations based on the W3C Web Annotation data model allowing the user to link the URL of a given item of content on the Web with additional information.

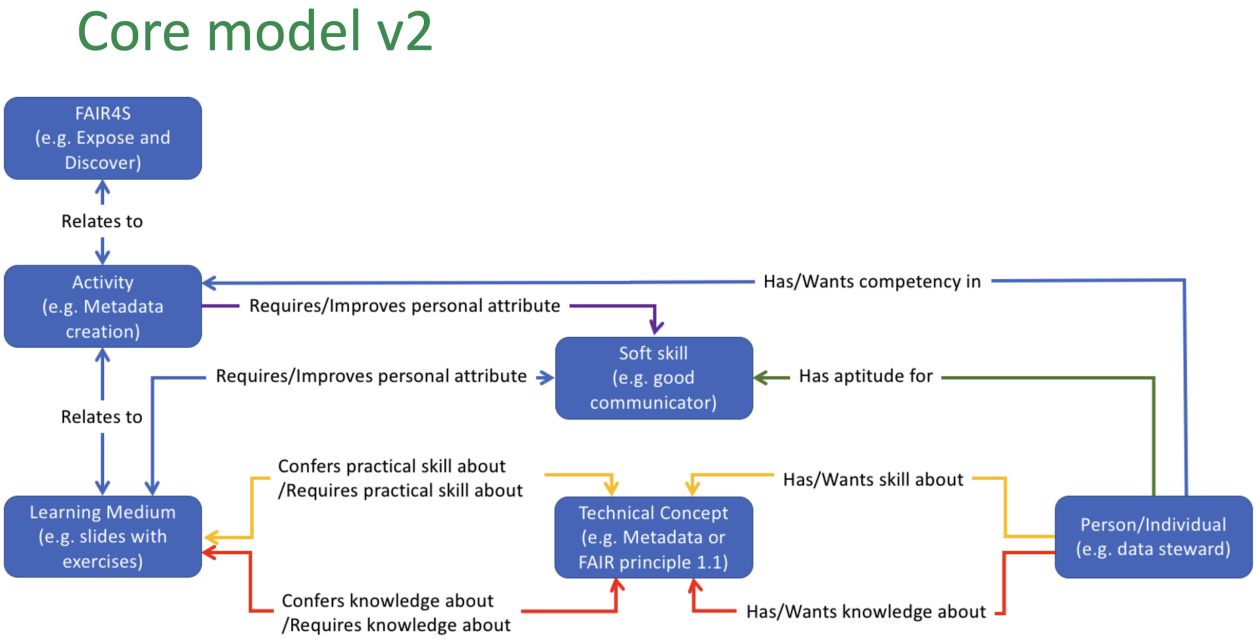

Figure 3: Core model v. 2.

In the current version, the service allows the user to create semantic annotations, free text keywords, comments and assertions (formatted as nanopublications). To support the annotation process with semantic concepts, the service offers an autocomplete function relying on an index of the concepts of terms4FAIRSkills. This tool has been used during dedicated hackathons during which we first trained the participants on its usage and on the annotation process. As mentioned above, we organised an initial pair of hackathons to test the annotation process and evaluate any potential changes that could be made to Semaphora for supporting the annotator. In the first session, we used RightField for the annotations and worked on nine different training materials from the CODATA-RDA School of Research Data Science. During the second pair of hack sessions, domain experts used Semaphora and we collected 38 new concepts and 90 semantic annotations on twelve documents (seven Google documents and five Zenodo records).

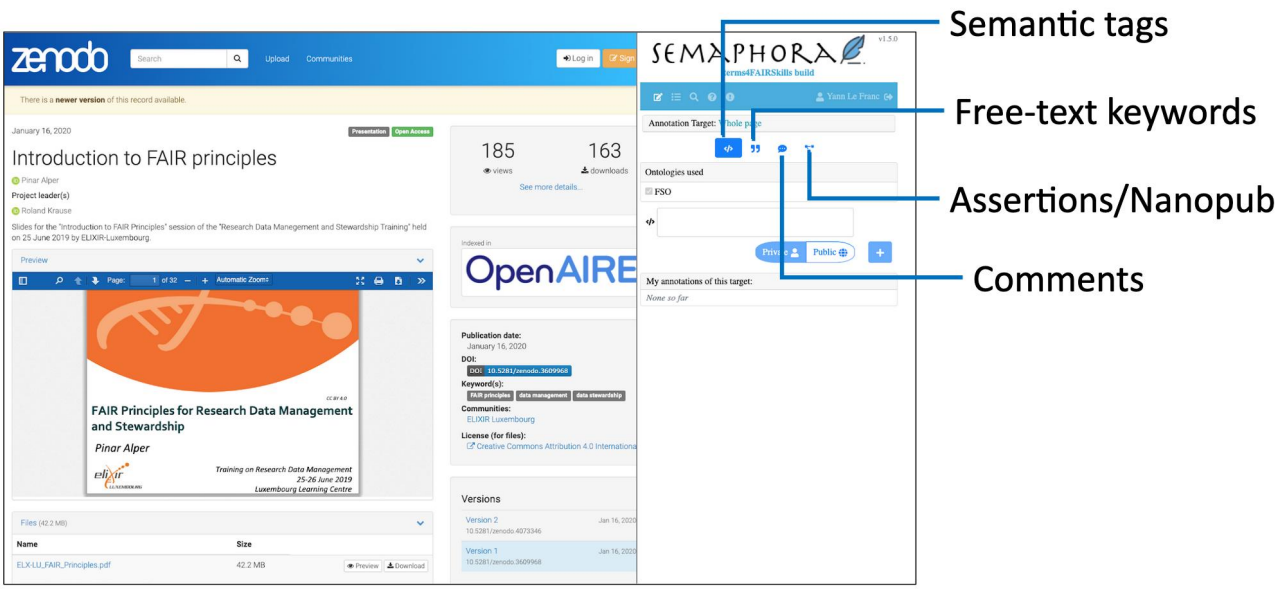

Figure 4: The Semaphora plugin, annotating an online training resource from ELIXIR.

Through further hackathons and subsequent feedback, we continued to refine the terminology. This fifth stage resulted in a reduction in the number of terms and an increase in definitions, synonyms and relationships as we increased the complexity of the terminology and moved away from a flat list of terms without synonyms. Feedback also indicated that featuring the FAIR4S class separately from the activity class was confusing for users and indicated a historical rather than practical separation. Further refinement of the technical concepts class resulted in more terms falling within the FAIR principles hierarchy, which allows for more efficient searching for directly-applicable FAIR concepts. These changes resulted in an adjusted model (v3 - figure 5) and an overall reduction in classes from 746 to 605, with 262 classes imported from the CASRAI RDM Glossary.

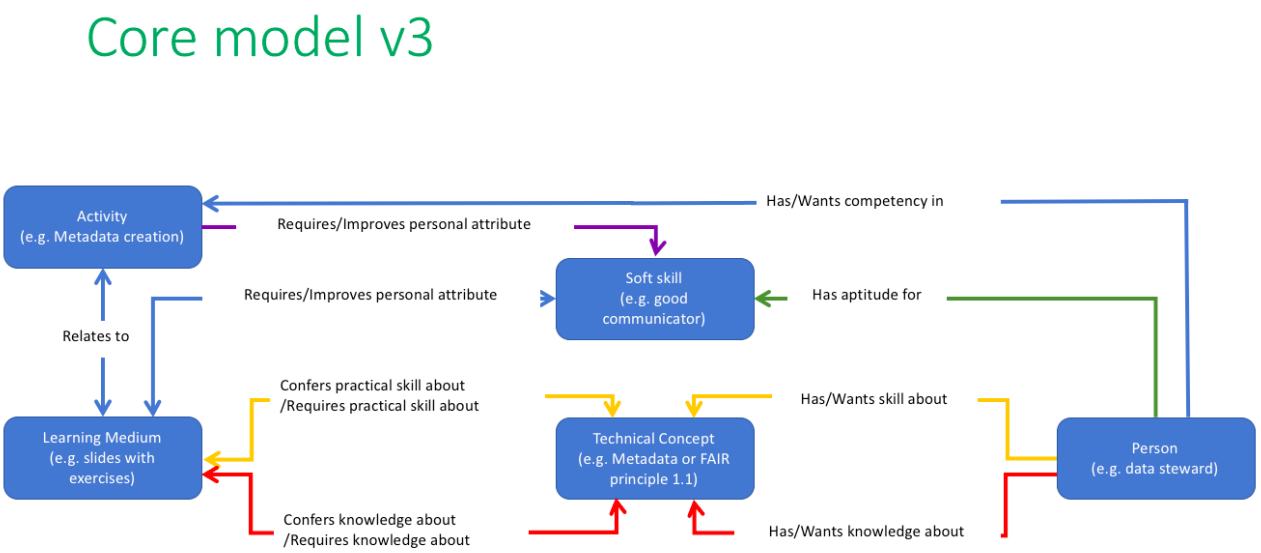

Figure 5: Core model v. 3.

We continued to refine the terminology, based on the annotations and feedback received from the final “Annotate-a-thon” hack week of annotation activity. In parallel, annotated training material can be retrieved through the Semaphora search interface demonstrating how annotation can federate content from multiple sources.

Following Open Science and FAIR Semantics recommendations, the latest version of the terminology can always be found on the terms4FAIRskills Github repository, alongside older versions and an issue tracker, should users wish to suggest changes. We will continue to refine the terminology to serve our growing community and to ensure we can provide a cross-EOSC framework for the annotation and retrieval of FAIR and Open data stewardship courses and materials. Through this process, we have created a framework of FAIR KSAs, relating technical concepts and transferable skills to the roles that need them, and the learning media that can confer these competencies, such that the user can perform activities to ensure their data, code and indeed any research outputs are FAIR and as Open as possible.

3.2. Additional outputs and outcomes

The funded activity also delivered the following outcomes:

- Growth in the community surrounding the terminology;

- Increased understanding in that community of the curation process;
- Proof of concept use of Semaphora, derived from B2Note, which has an open source version and is itself an output of EOSC funding; likely to be widely used;
- Benefit of familiarization of the user community with the use of the terminology, curation process and the Semaphora annotation tool.

The community-building around the development of the terminology is critical. The terminology is both serving and being shaped by a range of stakeholders who participated in our events, providing verbal and written feedback throughout the funded phase of activity (for example, see the mini user stories in the Appendices). By carefully listening to our community, we have developed and improved the terminology and have raised awareness of both the value of concretising FAIR skills into a structured terminology and the value of annotating existing training materials with this terminology to increase their potential findability and reusability. In addition to the participatory events, awareness and communication activities were also carried out via blogposts and Twitter (see section 4. 'Publicly available outputs').

We have also increased understanding across this community of the Semaphora tool and its use to interrogate the terminology and apply annotations to any online digital object. In turn, exposing the Semaphora tool to this sort of group testing generated valuable feedback and initiated further development and improvement of this tool.

We asked some of our most active community members to reflect on their participation immediately after the end of the project. These quotes demonstrate the enthusiasm of the participants for the work of this funded activity.

Dr Celia van Gelder, co-lead of the ELIXIR Training Platform, commented: "*Terms4FAIRskills is a great bottom-up initiative that has now run for several years and ELIXIR Training has been involved from the beginning in the Coordination Group. For me, this is an excellent example of representatives of different communities working slowly but surely together to achieve something that will be very valuable for all. The training community needs a vocabulary to annotate their FAIR-related training materials consistently. With the Semaphora tool an impressive step forward has been accomplished; Semaphora is very easy and intuitive to use by the annotators, which is very important. For the next step we can now move forward and open up to a larger group of annotators, and together enrich and grow the vocabulary and start implementing it.*"

Dr Angus Whyte, Senior Institutional Support Officer at the Digital Curation Centre, reflected: "*This felt like a very useful session to me. Having the Semaphora tool made quite a difference. It was quite helpful that it hides the hierarchical structure in the terminology, as we know that is a work-in-progress. The value of the work that has gone into distilling the terms was pretty clear, as was the ease of using the tool. A bit of practice with the annotation helped too. It takes a bit of discipline to be economical with the choice of terms to tag an item with. And we had an interesting and useful discussion about the level of granularity at which the annotation is most helpful.*"

Prof Hugh Shanahan, Professor of Open Science at Royal Holloway, University of London, agreed: "*The [hack] sessions were enormously helpful in understanding what's been achieved because of the EOSC [Co-creation] funding. The project has reached a significant level of maturity and it's much clearer now what the finished project will look like as a tool for annotating and finding FAIR-related training materials.*"

3.3 Summary and next steps

Thanks to the EOSC Co-creation fund, we have been able to take a number of important steps forward. Firstly, through increased and repeated interaction with the community, we have been able to grow and refine the terminology itself. The terminology now has a greater number of terms, definitions and relationships. More importantly, the terminology model has been refined to better serve the two use cases and to provide richer linked annotation. Terms4FAIRskills has also provided an excellent use case as a pilot for the Semaphora annotation tool, which has benefited from use in the hack sessions and has also provided a suitable environment for increased and more accurate annotation.

Our funded period was extremely busy with respect to the technical development of the terminology and the Semaphora tool and with community building and liaison activities. We will continue to refine and develop the terminology based on the rich feedback received during our program of meetings and hack sessions. Our next goal is to deploy a refined, stable version of the terminology in an applied context, for example for the annotation of life science training material in TeSS, or in the FAIRsFAIR competence centre.

4. List of publicly-available outputs

The project has produced a number of publicly available outputs during its funded period.

- Website: https://terms4fairskills.github.io/
- Latest version of the terminology: https://github.com/terms4fairskills/FAIRterminology
- 22nd October 2020: EOSC Symposium. Presentation slides: https://docs.google.com/presentation/d/1mMjMO1mdl1UyL4_X43Tv4wqlJBZnf_YC/edit#slide=id.p1
- 30 November and 2 December 2020 (2 iterations): GO-FAIR/CODATA FAIR Convergence Symposium, panel on 'FAIR Data Stewardship: training and career opportunities'. Presentation video: https://drive.google.com/file/d/10GGnltlKJtCTpaYo5-9jqPlHUXFhYigf/view?usp=sharing; panel webpage: https://conference.codata.org/FAIRconvergence2020/sessions/259/
- 12th February 2021: ELIXIR Converge. Presentation slides: https://drive.google.com/file/d/1jwsuZZGHANpLBsiFPydZ9DuM6-pwHT6U/view?usp=sharing
- 22nd February 2021: ELIXIR EDAM and FAIR terminology. Presentation slides: https://drive.google.com/file/d/1Vj1-eE6PfbhmbQmU3uFLPxSPQjneI8hq/view?usp=sharing
- 28th February 2021: ELIXIR FAIR Training Working Group. Presentation slides: https://docs.google.com/presentation/d/16EsxMyNCh9hEFJJOIS2Dxq8ma5_jDCGZ/edit#slide=id.p1
- Blogposts:

- CODATA blog: https://codata.org/blog/category/terms4fairskills/;
- FAIRsharing blog: https://blog.fairsharing.org/?p=89;
- FAIRsFAIR project news page: https://www.fairsfair.eu/articles-publications/two-more-terms4fairskills-hackathons-start-new-year; https://www.fairsfair.eu/articles-publications/fair-data-terms-refined-terms4fairskills-hackathon

Appendix 1: Personnel

The EOSC Co-creation fund resourced a small core team from 22 Sep 2020 to 22 Feb 2021. This core team consisted of Yann Le Franc, Pete McQuilton, Laura Molloy, Simon Hodson and Sally Davies.

In addition, several stakeholders made in-kind contributions to the project: Hugh Shanahan, Royal Holloway, UK; Celia van Gelder, DTL, ELIXIR-NL, NL; Susanna Assunta Sansone, University of Oxford; Angus White, DCC, UK; Kevin Ashley, DCC, UK.

Terms4FAIRskills hack sessions, 22/9/2020 - 26/3/2021: participants

In addition to the core project team, the following individuals took part in our participatory events:

Victoria Dominguez del Angel, ELIXIR / Institut Français de Bioinformatique

Marjan Grootveld, DANS

Erik Hjerde, Norges Arktiske Universitet

Ellen Leenarts, DANS

Allyson Lister, FAIRsharing, University of Oxford

Marie-Anne Maurel, CINES

Elizabeth Newbold, UKRI STFC

Paula Oset Garcia, Universiteit Gent

Olivier Sand, ELIXIR

Helena Schnitzer, Bielefeld University

Hugh Shanahan, Royal Holloway, University of London

Celia van Gelder, ELIXIR / DTL

Shanmugasundaram Venkataraman, Digital Curation Centre

Louis Vinchon, CINES

Samuel Viscapi, CINES

Angus Whyte, Digital Curation Centre

Appendix 2: Terms4FAIRskills user story: ELIXIR Training

ELIXIR is the international distributed research infrastructure for life sciences data and training is coordinated via the Training Platform (TrP) (https://elixir-europe.org/platforms/training), one of five technical platforms or areas of work within ELIXIR. The mission of the TrP is to strengthen national training programmes, grow bioinformatics training capacity and competence across Europe, and empower researchers to use ELIXIR's services and tools. Since 2020, within the ELIXIR-CONVERGE project (https://elixir-europe.org/about-us/how-funded/eu-projects/converge/wp2), training in data management/data stewardship has become an additional focus of the TrP.

The TrP consists of both a technical and a human infrastructure (the Training Coordinators Group TrCG). The ELIXIR Training Portal TeSS (https://tess.elixir-europe.org) is a registry for training materials and events. TeSS enables training providers, from within but also from outside ELIXIR, to register training events and materials, facilitating the attribution of credit and ownership to their authors.

The ELIXIR TrP has been involved in the terms4FAIRskills coordination team since the very beginning, when it resulted from the GO-TRAIN initiative. Terms4FAIRskills set out to develop a community defined and endorsed terminology to annotate training materials, events, learning paths, tools etc. Such a terminology is needed to make training resources (more) discoverable and (re)usable. Although the training materials from the TrP (which are registered in TeSS) are already annotated with the EDAM terminology, it became clear that an additional vocabulary was needed that could complement EDAM, especially for the materials and courses related to Data Management and FAIR Data Stewardship. For this reason the ELIXIR Training Platform is actively exploring and testing terms4FAIRskills and we expect that the terminology will meet our needs.

Contact person for ELIXIR Training Platform: Celia van Gelder (Training Platform co-lead, celia.van.gelder@dtls.nl)

Appendix 3: Terms4FAIRskills user story: CODATA-RDA Schools of Research Data Science

The CODATA-RDA Schools of Research Data Science originated from a collaborative effort between CODATA and the Research Data Alliance to produce a series of schools to provide training to early career researchers (ECRs) in the necessary research data science skills to make them ready for the data revolution. The focus of delivery is for ECRs from low and middle income countries. This training covers a wide range of topics including research data management. and open and responsible research. The FAIR principles represent an important component of the training.

Correspondingly, the FAIRsFAIR project addresses the development and concrete realisation of an overall knowledge infrastructure on academic quality data management, procedures, standards, metrics and related matters, based on the FAIR principles.

The Schools have partnered with the FAIRsFAIR project to further develop the ECR training and to provide instructor training to data stewards. As a result, a new set of materials specifically for Data Stewards has been developed. The Data Steward materials will be made available in a competence centre created by the FAIRsFAIR project to promote findability, accessibility and reuse of training materials for FAIR data skills.

terms4FAIRskills has been used to annotate the substantial range of materials that has been generated by the Schools and FAIRsFAIR, and the Schools has been a partner of the terms4FAIRskills project since its start. The terminology will improve the findability of our materials for both ECRs and Data Stewards and, as we provide more training via an online approach, will be key in encouraging their uptake.

Contact person for CODATA-RDA Schools of Research Data Science: Hugh Shanahan (Hugh.Shanahan@rhul.ac.uk)

Appendix 4: Terms4FAIRskills user story: EOSC-Pillar RDM Training and Support Catalogue

EOSC-Pillar is an INFRAEOSC-05 project involving 18 European partners, whose ultimate purpose is to support the implementation of the EOSC by leveraging national initiatives in Austria, Belgium, France, Germany and Italy and thematic initiatives developed by research communities working in national and European collaborations.

Within the project, a specific task focuses on Research Data Management (RDM) standards and on creating awareness in different data stewardship approaches in order to harmonize support offered to researchers. In the framework of this task, an RDM Training and Support catalogue has been published to allow the diffusion and adoption of FAIR principles in research data management. The catalogue contains training materials, but it also includes day-to-day, operational and readily available resources that can be used by data stewards to support researchers.

To describe the items in the catalogue, specifications for metadata elements and values were adopted based on recent outputs from the WP6 (FAIR Competence Centre) of the FAIRsFAIR project and the FAIR4S skills and capability framework from EOSC Pilot. Using the terms4FAIRskills terminology will increase the findability of the training and support materials in the catalogue, and improve the precision of annotations with regards to the skills and competences that the materials help to develop.

Although the EOSC-Pillar RDM Training and Support catalogue joined the terms4FAIRskills initiative at a later stage, we will continue to use the terminology developed to annotate the resources in the catalogue and provide feedback and interact with the terms4FAIRskills team as necessary.

Contact person for EOSC-Pillar RDM Training and Support Catalogue: Paula Oset-Garcia (Paula.OsetGarcia@UGent.be)