Open Methodology

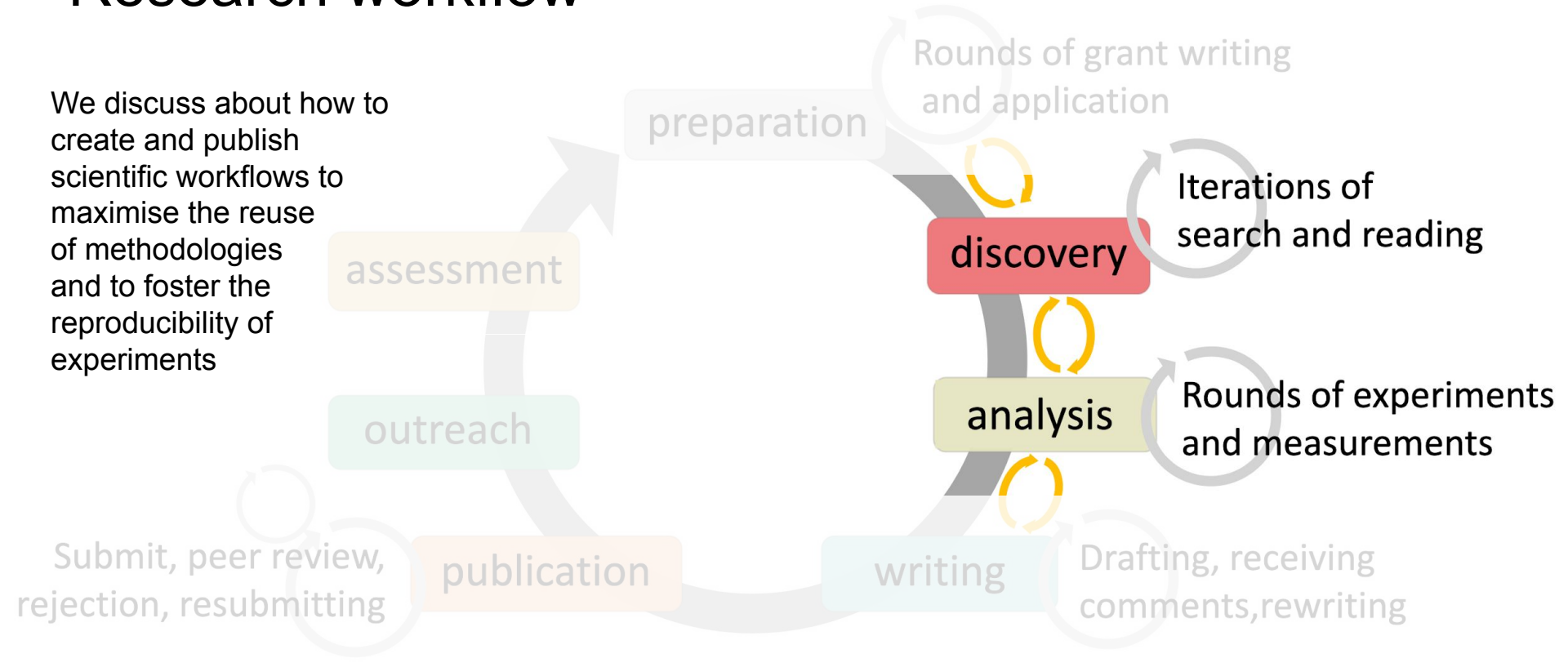
Silvio Peroni
silvio.peroni@unibo.it – https://orcid.org/0000-0003-0530-4305 – @essepuntato

Open Science (A.Y. 2022/2023)
Second Cycle Degree in Digital Humanities and Digital Knowledge
Alma Mater Studiorum - Università di Bologna

DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA

Research workflow

We discuss about how to create and publish scientific workflows to maximise the reuse of methodologies and to foster the reproducibility of experiments



Kramer, B., & Bosman, J. (2015, June 18). The good, the efficient and the open—Changing research workflows and the need to move from Open Access to Open Science. CERN Workshop on Innovations in Scholarly Communication (OAI9), University of Geneva, Geneva, Switzerland. https://www.slideshare.net/bmkramer/the-good-the-efficient-and-the-open-oai9

What is a methodology

A methodology is the definition and description of the methods used in a study to collect and analyse the data necessary to answer to specific research questions

In research articles, usually it is introduced within the “Methods” section, which “should provide the readers with sufficient detail about the study methods to be able to reproduce the study if so desired”

Such a section “should be specific, concrete, technical, and fairly detailed” and should contain information about “[t]he study setting, the sampling strategy used, instruments, data collection methods, and analysis strategies” adopted to answer the research questions

Perneger, T. V. (2004). Writing a research article: Advice to beginners. International Journal for Quality in Health Care, 16(3), 191–192. https://doi.org/10.1093/intqhc/mzh053

Is the “Methods” section enough?

Often, the “Methods” section does not provide enough information to clearly reproduce the setting and the experiment

Some reasons:

- Word limit (still used today) of research articles depending on the journal
- Lack of transparency in the procedure followed for the experiment

from https://doi.org/10.1007/s11192-020-03397-6,
freely available at https://arxiv.org/abs/1903.06142

Methods and material

The initial bibliographic and citation data we used in our analysis have been gathered and analysed through Semantic Publishing technologies. Semantic Publishing (Shotton 2009) concerns the use of Web and Semantic Web technologies and standards for enhancing a scholarly work (e.g. using plain RDF statements Cyganiak et al. 2014) to improve its discoverability, interactivity, openness and (re-)usability for both humans and machines. The assumptions of openness implicit in Semantic Publishing have been explicitly adopted for the publication of research data by the FAIR (Findable, Accessible, Interoperable, Re-usable) data principles (Wilkinson et al. 2016). Early examples of the semantic enrichment of scholarly works involved the use of manual (e.g. see Shotton et al. 2009) or (semi-)automatic post-publication processes (e.g. see Peroni 2017).

Semantic Publishing technologies allow one to enrich the semantic payload of the networks of published articles, usually linked through their plain citation links, in order to describe several content-related and context-related aspects of the publishing domain. It would be possible to group such semantic enrichment in eight different buckets. Each bucket can be able to describe a particular semantic specification of an article. For instance, it can concern either the description of the article content from different angles (e.g. structure, rhetoric, argumentation) or contextual elements relating to the creation of a paper (e.g. research project, people contributions, publication venue) (Peroni 2017).

What is an (open) methodology

An open methodology is a methodology described in sufficient detail to allow other researchers to repeat the work and apply it elsewhere

It is not always possible, of course, to access the same unique resource of the original study, e.g. a specific computer hardware

However, it is important to clearly describe the full methodology because it is the only way to enable reproducibility and to allow others to learn from what scientists have done

Crucial aspect: to reveal how a scientist carried out an experiment is at the heart of any study

Watson, M. (2015). When will ‘open science’ become simply ‘science’? Genome Biology, 16(1), 101. https://doi.org/10.1186/s13059-015-0669-2

Reusing existing research

The “discovery” step is an important aspect related to the definition of the methodology one wants to define to address research questions

It is possible that prior works have addressed either the same or similar research questions, or that the methodology proposed in a particular research can be reused to address, in principle, different research questions

Thus, it is crucial to perform a literature review to understand whether prior work can be of any help – following the principle of standing on the shoulders of Giants

A literature review should provide a concise examination and discussion of evidence in a particular area and is usually present in several written research outputs such as proposals for funding, proposals for academic degrees, research articles, guidelines for professional practice, and reports

Bolderston, A. (2008). Writing an Effective Literature Review. Journal of Medical Imaging and Radiation Sciences, 39(2), 86–92. https://doi.org/10.1016/j.jmir.2008.04.009

Literature review

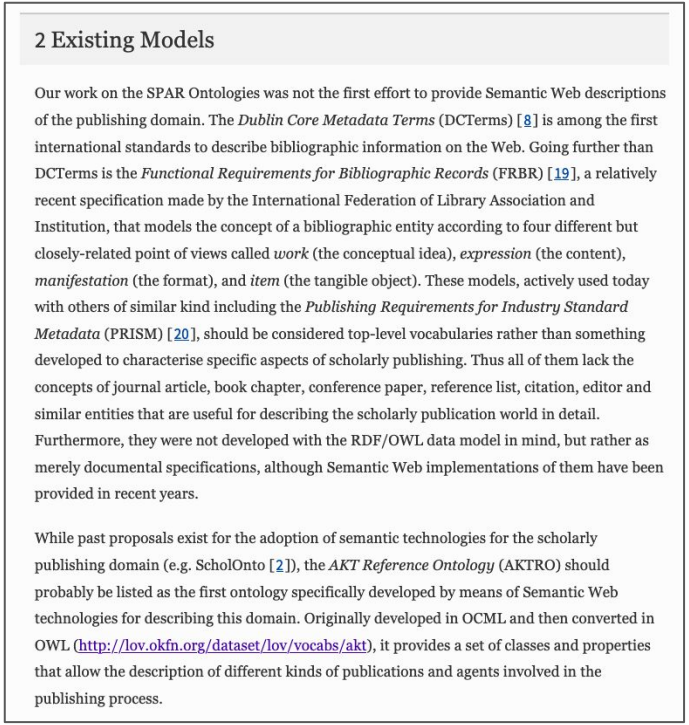
A literature review should be:

- Comprehensive, gathered from all relevant sources
- Fully referenced, allowing others to follow the author's argumentation
- Selective, adopting search strategies to find key works
- Relevant, focusing on pertinent data
- Balanced, including different ideas and opinions
- Critical, assessing existing works
- Analytical, providing new points of view on a topic

Main type of reviews:

- Systematic review: an attempt to quantitatively condense the results from several articles into a single statistic
- Introduction to a primary research topic: used to set the scene for a primary research topic, essential to introduce a study, its methods, and to provide a foundation for the discussion of the results

from https://doi.org/10.1007/978-3-030-00668-6_8



2 Existing Models

Our work on the SPAR Ontologies was not the first effort to provide Semantic Web descriptions of the publishing domain. The *Dublin Core Metadata Terms* (DCTerms) [8] is among the first international standards to describe bibliographic information on the Web. Going further than DCTerms is the *Functional Requirements for Bibliographic Records* (FRBR) [19], a relatively recent specification made by the International Federation of Library Association and Institution, that models the concept of a bibliographic entity according to four different but closely-related point of views called *work* (the conceptual idea), *expression* (the content), *manifestation* (the format), and *item* (the tangible object). These models, actively used today with others of similar kind including the *Publishing Requirements for Industry Standard Metadata* (PRISM) [20], should be considered top-level vocabularies rather than something developed to characterise specific aspects of scholarly publishing. Thus all of them lack the concepts of journal article, book chapter, conference paper, reference list, citation, editor and similar entities that are useful for describing the scholarly publication world in detail. Furthermore, they were not developed with the RDF/OWL data model in mind, but rather as merely documental specifications, although Semantic Web implementations of them have been provided in recent years.

While past proposals exist for the adoption of semantic technologies for the scholarly publishing domain (e.g. ScholOnto [2]), the *AKT Reference Ontology* (AKTRO) should probably be listed as the first ontology specifically developed by means of Semantic Web technologies for describing this domain. Originally developed in OCML and then converted in OWL (http://lov.okfn.org/dataset/lov/vocabs/akt), it provides a set of classes and properties that allow the description of different kinds of publications and agents involved in the publishing process.

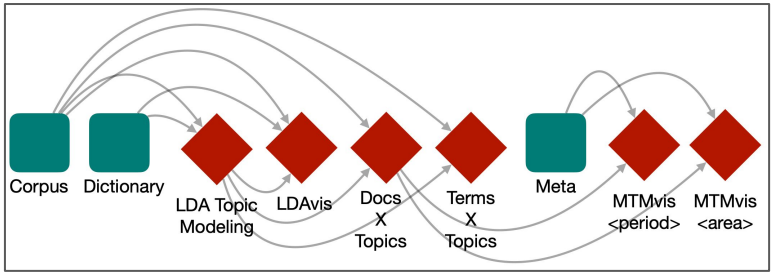
Bolderston, A. (2008). Writing an Effective Literature Review. Journal of Medical Imaging and Radiation Sciences, 39(2), 86–92. https://doi.org/10.1016/j.jmir.2008.04.009

Workflow

A workflow is a way to define an experiment (and the methods characterising it) as a directed graph where the nodes represent operations and edges specify dependencies between the operations

Once specified, a workflow can be reused by other scientists and enables the understanding an experimental process, the replication a previous experimental result, or its repurposing as a building-block in the design of new workflow-based experiments

Issue: workflows may decay, i.e. could not be understood, or executed when downloaded



Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K., Palma, R., Mina, E., Corcho, O., Gómez-Pérez, J. M., Bechhofer, S., Klyne, G., & Goble, C. (2015). Using a suite of ontologies for preserving workflow-centric research objects. Journal of Web Semantics, 32, 16–42. https://doi.org/10.1016/j.websem.2015.01.003

Crucial aspects for defining workflows

Example data inputs: make available exemplar input data accompanied by their description that enable the successful execution of the workflow, otherwise both experiment reproducibility and the ability to understand the function the workflow is inhibited

Long-term archivability: preserve the workflow in appropriate systems together with provenance traces of its data results, to track how results were produced by the workflow and to enable repairing broken workflows

Appropriate documentation: describe, in natural language, all the steps of the analysis, including its inputs, intermediate steps, and outputs, and make them annotable, since insufficient documentation impairs the runnability and understandability of workflows

Trackability of changes: use tools and approaches to identify changes of a workflow (including in the data and the environments used to run it), which is of particular importance when third party resources, that are not under the author's control, are used – the goal is to allow a user

- to retrieve the original version of input data and environment setting in order to reproduce/verify the original results
- to retrieve the different parameter configurations used to generate the different versions of outputs
- to identify the changes made to the workflow specification

Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K., Palma, R., Mina, E., Corcho, O., Gómez-Pérez, J. M., Bechhofer, S., Klyne, G., & Goble, C. (2015). Using a suite of ontologies for preserving workflow-centric research objects. Journal of Web Semantics, 32, 16–42. https://doi.org/10.1016/j.websem.2015.01.003

Computational workflows

In computational workflows each task represents the execution of a computational process, such as running a piece of code, invocating a service, calling a tool, accessing to a database, submitting a job to a compute cloud, executing data processing script, etc.

Computability is an added value to have for fostering the reproducibility of workflows

Python notebooks can be used to create reproducible computational workflows that can be run within the Jupyter environment

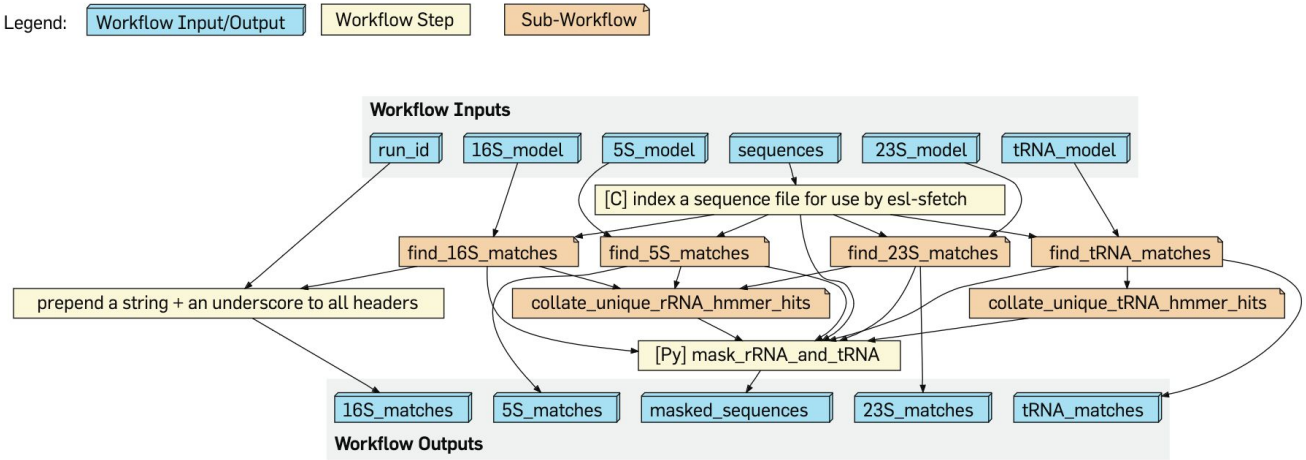
There are several tutorials that demonstrate how to use Jupyter for this activity, such as the one made available by Justin Kitzes, Shane Grigsby, Mark Wilber of the University of California and Santa Barbara

Beg, M., Taka, J., Kluyver, T., Konovalov, A., Ragan-Kelley, M., Thiery, N. M., & Fangohr, H. (2021). Using Jupyter for Reproducible Scientific Workflows. Computing in Science & Engineering, 23(2), 36–46. https://doi.org/10.1109/MCSE.2021.3052101, freely available at https://arxiv.org/abs/2102.09562

Common Workflow Language

The Common Workflow Language (CWL) is a set of open standards for describing and sharing computational workflows

It support the specification of critical concepts such as automation, scalability, abstraction, provenance, portability, and reusability, and it has been developed around core principles of community and shared decision making, reuse, and zero cost for participants



Crusoe, M. R., Abeln, S., Iosup, A., Amstutz, P., Chilton, J., Tijanić, N., Ménager, H., Soiland-Reyes, S., Gavrilović, B., Goble, C., & Community, T. C. (2022). Methods included: Standardizing computational reuse and portability with the Common Workflow Language. Communications of the ACM, 65(6), 54–63. https://doi.org/10.1145/3486897

The ideal: FAIR computational workflows

Properly designed workflows contribute to make data FAIR, since they provide the metadata and provenance necessary to describe their data products and they describe the involved data in a formalized, completely traceable way

In addition, workflows are research products in their own right, encapsulating methodological know-how that is to be found and published, accessed and cited, exchanged and combined with others, and reused as well as adapted – and, thus, they should be compliant with the FAIR principles

Need to extend somehow FAIR indicators to address the processual nature of workflows – and it is a key component to have since a framework for FAIR workflows will enhance reproducibility, quality and transparency of the data generated, but also of the processing path that lead to the data results

Goble, C., Cohen-Boulakia, S., Soiland-Reyes, S., Garijo, D., Gil, Y., Crusoe, M. R., Peters, K., & Schober, D. (2020). FAIR Computational Workflows. Data Intelligence, 2(1–2), 108–121. https://doi.org/10.1162/dint_a_00033

The reality: experiments in defining FAIR workflows

Studies tried to apply a FAIRification process, i.e. transforming an existing workflow to its FAIR version, by leveraging the RDF technology for the interoperability aspect

Example: FAIRify the PREDICT workflow, a workflow based on machine learning

Approach:

1. Creation of a unified model that reuses several semantic models to show how a workflow can be semantically modeled
2. Publication of the workflow representation, data and metadata in a RDF triplestore which was used as FAIR data point
3. Definition of competency questions (e.g. “What are the existing versions of a workflow and what are their provenances?”) that can be answered through SPARQL queries

Issues identified: reusing of existing semantic models is challenging task, in particular when they present reproducibility issues, different conceptualizations, and overlapping terminology

Celebi, R., Rebelo Moreira, J., Hassan, A. A., Ayyar, S., Ridder, L., Kuhn, T., & Dumontier, M. (2020). Towards FAIR protocols and workflows: The OpenPREDICT use case. PeerJ Computer Science, 6, e281. https://doi.org/10.7717/peerj-cs.281

Registered reports

Recently, a new approach to journal publishing has been introduced in order to foster the definition of better methodologies in research environments, i.e registered reports

Registered reports can be seen as a publication of the background information and the methodology defined for conducting a research before the results obtained by running such a methodology are either produced or revealed

In practice:

1. The authors submit a registered report to a journal
2. The report is evaluated via the usual peer-review process, to obtain constructive criticisms and recommendations by peers (i.e. the reviewers) to improve the protocol proposed
3. The revised report is provisionally accepted and published in the journal as an article
4. The authors perform the experiment and expand the journal article to include both the results and discussion
5. The new part of the article is again peer-reviewed but, providing that the authors have implemented the agreed protocol, publication should be guaranteed

Benefits: minimising the potential for publication bias when publishing negative vs positive results and checking the soundness of the methodology before running the experiment

Clarke, P., Buckell, J., & Barnett, A. (2020). Registered Reports: Time to Radically Rethink Peer Review in Health Economics. PharmacoEconomics - Open, 4(1), 1–4. https://doi.org/10.1007/s41669-019-00190-x

Adoption of registered reports

Several journals has started to adopt registered reports as a standard practice for research – e.g. PLOS One and Royal Society Open Science

from https://everyone.plos.org/2020/01/14/registered-reports-are-coming-to-plos-one/

The first stage is the **Registered Report Protocol**. This new publication type describes the proposed rationale, methodology, and any ethical approvals needed for the work. We'll peer review this initial phase of research to ensure the study's scientific rigor and that the planned research will meet *PLOS ONE*'s publication criteria. If accepted, authors proceed to their investigation with the promise that their subsequent work describing the full study and all of its findings, will be accepted and published by *PLOS ONE* as a **linked Research Article**, provided that the authors adhere to the initial study design and conduct experiments to *PLOS ONE*'s standards of rigor.

A Registered Report is a form of journal article in which methods and proposed analyses are pre-registered and peer-reviewed prior to research being conducted (stage 1). High quality protocols are then provisionally accepted for publication before data collection commences. The format is open to attempts of replication as well as novel studies. Once the study is completed, the author will finish the article including results and discussion sections (stage 2). This will be appraised by the reviewers, and provided necessary conditions are met, **publication is virtually guaranteed**.

The main benefits of this two-stage approach are:

- Once the methods and proposed analyses are provisionally accepted, the journal will commit to publishing the results regardless of the outcome, provided the final study conforms to the initially approved proposal and meets all quality checks. This means that publication bias is reduced as negative results will not prevent publication.
- Peer review of the research proposal provides an opportunity for the authors to receive constructive critical feedback that may help them to fine-tune the study design prior to conducting the experiment.
- This process can help reduce researcher bias.
- This process may enhance the credibility of the work.

from https://royalsocietypublishing.org/rsos/registered-reports

Pure workflow publications

In order to support increased sharing of open research methodologies, some journals (e.g. PLOS One and MethodsX) have started to publish directly methodology articles as part of their portfolio without having them explicitly linked with the results of an experiment

Some justification for this choice:

- methods development and sharing deserve increased recognition, via a peer-reviewed publication
- understanding the exquisite details that make methods work is essential for reproducibility and for accelerating science
- researchers who develop methods deserve more recognition

Hrynaszkiewicz, I. (2020, December 7). Show your work. Peer-Reviewed Protocols. The Official PLOS Blog. https://theplosblog.plos.org/2020/12/show-your-work-peer-reviewed-protocols/

End

Open Methodology

Silvio Peroni
silvio.peroni@unibo.it – https://orcid.org/0000-0003-0530-4305 – @essepuntato

Open Science (A.Y. 2022/2023)
Second Cycle Degree in Digital Humanities and Digital Knowledge
Alma Mater Studiorum - Università di Bologna

DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA