Journal of the Text Encoding Initiative

Issue 14 | 2021
Selected Papers from the 2019 TEI Conference

Archiving a TEI Project FAIRly

Andrew Creamer, Gaia Lembi, Elli Mylonas and Michael Satlow

Electronic version
URL: https://journals.openedition.org/jtei/4324
DOI: 10.4000/jtei.4324
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Andrew Creamer, Gaia Lembi, Elli Mylonas and Michael Satlow, "Archiving a TEI Project FAIRly", *Journal of the Text Encoding Initiative* [Online], Issue 14 | 2021, Online since 08 December 2022, connection on 04 February 2023. URL: http://journals.openedition.org/jtei/4324 ; DOI: https://doi.org/10.4000/jtei.4324

Archiving a TEI Project FAIRly

Andrew Creamer, Gaia Lembi, Elli Mylonas, and Michael Satlow

ABSTRACT

The Inscriptions of Israel/Palestine project is an online corpus of over four thousand inscriptions from Israel and Palestine, written in Hebrew, Greek, Latin, and Aramaic, dating roughly from the Persian Period to the Arab Conquest. The source files with inscription text and metadata are encoded using EpiDoc, a TEI customization widely used by epigraphers. As the project prepared to deposit its XML files in an institutional repository, it transformed them into a locally developed robust archival format. This paper evaluates these decisions against the FAIR metrics, using IIP as a test case. This allows us to suggest improvements for our own archival encoding as well as to see where EpiDoc and TEI enhance FAIRness and where they could provide more support. Finally, we suggest some ways to use FAIR metrics that are more amenable to TEI documents and corpora.

INDEX

Keywords: FAIR, digital epigraphy, EpiDoc

1. Introduction to IIP, history and motivation, principles

1 The Inscriptions of Israel/Palestine project (IIP) collects and transcribes inscriptions from Israel/Palestine, dating roughly from the Persian period through the Islamic conquest, and makes them discoverable and browsable on the web. There are about ten thousand relevant texts, written in Hebrew, Aramaic, Greek, and Latin, by Jews, Christians, Greeks, and Romans. Their number and variety provide a fascinating window into the ancient world. Despite their importance for historical and linguistic investigation, these inscriptions have been published for the most part individually or in small corpora; IIP provides an accessible platform for searching and analyzing them all together.

2 Currently, the project contains over four thousand inscriptions, which are available as a searchable online corpus, via an API, and in the IIP GitHub repository.[1] The project team is continuing to add to the corpus, and at the same time developing new functionalities for the website and experimenting with interfaces for different audiences. The project is focusing its new developments on the inscription text and, in particular, using NLP for lemmatization and other lexical features.

3 The epigraphic texts in IIP have been extensively marked up using the EpiDoc schema (Elliott et al. 2021). As with many other digital corpora, IIP simultaneously balances two criteria: inscriptions must be encoded so as to facilitate specific online presentations, interfaces, and interactivity, and must at the same time use data structures and descriptors that result in a sustainable epigraphic archive that will last across presentation formats and can be used for different purposes. Our digital corpus thus adds value, not just by allowing for accessibility but also by adding contextual and linguistic information and interoperability with similar digital projects that use EpiDoc and other structured formats.

4 For example, IIP disambiguates geographic names and dates in the edition texts by linking to external authorities using linked open data (LOD) Uniform Resource Identifiers (URIs). Geographic locations refer to the Pleiades gazetteer (Bagnall et al. 2006–). Working with Pleiades was mutually beneficial: as we linked our material to the Gazetteer, we also contributed new places to it. IIP relies on Periodo (Rabinowitz, Shaw, et al., n.d.) to enrich dating information, linking temporal

information to existing periods. The incorporation of LOD references into IIP metadata provides functionality beyond disambiguation: IIP data can now be aggregated with any other data set that uses the same linked-data URIs to express time and location.

5 Although the epigraphic community is working on developing ontologies and authority lists for metadata, these are not yet standardized. IIP adds links to the Getty Art & Architecture Thesaurus (AAT) (Getty Research Institute, n.d.) to object types, but implements project-level authority lists for genres and materials. This makes encoding much easier and less prone to error, as it provides encoders with values to choose from. The closed list of values also enables faceted searching and makes adding a new value a more considered decision.

6 IIP is constantly in motion, with new inscriptions being added and others corrected. As such, the project team recently began to assess the strategy and workflow necessary for data preservation. IIP data is routinely made accessible on GitHub and from the project website, via an API, but still requires a more sustainable and secure solution. The first choice for preservation is a local one, using the Brown Digital Repository (BDR).[2] In order to store robust, reusable resources in the BDR, we developed an archival format for the inscriptions that ensured each file was self-documenting and did not refer to locally maintained external files. A major challenge was incorporating full bibliographical information, which is stored in Zotero. It was also necessary to add documentation about accessibility in the TEI Header and include essential metadata such as the principal investigator and contributors (with ORCID [Open Researcher and Contributor ID] persistent identifiers when available); rights statements; edition information; information on how to cite the inscription; and snapshots of internal authority lists for any references in the file.[3] External LOD references linked via URIs remain in the files.

7 As part of the archival process, a Digital Object Identifier (DOI) was minted using the DataCite Registration Agency (RA) for the Inscriptions of Israel/Palestine project, which resolves to the project home page. At the moment that is the only DOI; there is no separate DOI for IIP data or individual inscriptions. We discuss below how we might better use DOI identification.

8 This paper evaluates the strategy and workflow for archiving a living and connected corpus. We begin with IIP as the test case, and, as it is encoded in these formats, can extend our results to the EpiDoc and TEI Guidelines and schemas. We have sought to adhere to the principles of FAIR, making our inscriptions *findable*, *accessible*, *interoperable*, and *re-usable*. We will document how the IIP project has attempted to adhere to these goals and the challenges that we still confront.

2. FAIR Principles and FAIR Metrics

9 The seed of the term "FAIR" was originally planted in 2014 during a workshop in Leiden, the Netherlands, in which the participants produced a set of guiding principles and practices for helping to ensure the long-term discovery, access, and reuse of metadata and data by machines and people (Wilkinson et al. 2016, 3). Members of the FORCE11 community formed a working group to continue work on these principles and develop attribution practices that are both "human understandable and machine-actionable" (Data Citation Synthesis Group 2014). Wilkinson et al. (2016) published the FAIR Principles with the aim of ensuring that curators of digital objects —that is, "(meta)data" and their systems and practices for storing, formatting, describing, and dissemination—would be guided towards making them "machine-actionable": "enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals" (Wilkinson et al. 2016, 1).

10 According to the FAIR Principles, to be *findable*, digital objects "meta(data)" should be assigned globally unique and persistent identifiers, be described with metadata that capture the aforementioned identifier, and be indexed in a searchable resource. *Accessibility* is framed around the retrievability of metadata via open protocols, allowing for authentication and authorization (with the proviso that metadata should in all cases be accessible even in cases when the object they describe no longer is). The principles comprising *interoperability* focus on the utilization of standard programming languages, controlled vocabularies, and authoritative references. Lastly, *reusability* of digital objects relies on the capture of provenance and rights information, with these metadata formatted to a community/domain standard, necessary to provide potential users with the ability to both research veracity and provide attribution.

11 The European Union (EU) has been active in funding both infrastructure and initiatives that align with the FAIR Principles, as well as metrics for assessing their implementation. Notably, in 2016 the European Commission (EC) called for the establishment of what would become the European Open Science Cloud (EOSC). The following year the EC established its Expert Group on FAIR Data. In 2018 the group published its report "Turning FAIR into Reality" (European Commission, Directorate-General for Research and Innovation 2018), which laid out an action plan that included "embedded and sustained metrics." That same year, the EU's Research and Innovation program Horizon 2020 funded a project called FAIRsFAIR: Fostering FAIR Data Practices in Europe (FAIRsFAIR, n.d.). The UK and EU partners collaborating on the project will develop standards for FAIRsFAIR certification of repositories that will inform the Rules of Participation (RoP) and regulatory compliance for participation in the EOSC, which the EOSC governance structure will use "to establish whether components of the infrastructure function in a FAIR manner" (European Commission, Directorate-General for Research and Innovation 2016; 2018). In the United States, the embrace of FAIR by federal funders of scientific research is notable, particularly its inclusion in the text of recent funding opportunities announcements from the Department of Energy (DOE) and within the most recent strategic plan of the National Institute of Environmental Health Sciences (NIEHS), one of the National Institutes of Health (NIH).

12 The EU, UK, and US examples above may give the false impression that FAIR did not have an impact on researchers in the humanities and social sciences. On the contrary, since their debut the FAIR Principles have provided different domains and communities interested in the curation, dissemination, and preservation of knowledge in the form of digital objects with a framework in which to situate their systems and practices, and to assess their projects through the lens of FAIR. This resulted in many of these projects drafting "roadmaps" intended to align their projects with meeting certain markers with the context of FAIR Principles. In addition, the invocation of FAIR as an adjective to represent a project's inclusion of FAIRness among their goals has extended beyond the guidelines' emphasis on machines to now equally describe researchers' data handling practices. For example, in their proposal for a Cross-Linguistic Data Formats Initiative, Forkel et al. (2018, 2) describe the current data sharing and reuse practices of researchers in their field of linguistics as being "far away from being 'FAIR.'"

13 The publication of the FAIR Principles provided a lens through which scholarly communities can view their domain-specific data stewardship practices as well as the stewardship of other types of digital objects such as software (Lamprecht et al. 2020). For example, Koster and Woutersen-Windhouwer use the FAIR Principles as the foundation for creating their "FAIR Principles for Library, Archive and Museum (LAM) Collections" (Koster and Woutersen-Windhouwer 2018). The authors translate the broad FAIR principles into ones that specify identifiers, standards, rights, and provenance practices recognized as best practices within LAM communities. They suggest a "roadmap" for LAMs to achieve the goal of "FAIRness." While at that time they did not go as far as defining the metrics for assessing the degree of a collection's FAIRness, their roadmap leaves open the possibility for this by recommending LAMs create working groups to study their collections' alignment with the FAIR principles and develop a FAIR policy and implementation plan. Other scholarly communities similarly moved from using the Principles as guide to using them as a tool for assessing the FAIRness of their systems and practices for describing and disseminating digital objects in their field. For example, Calamai and Frontini (2018) relied on the FAIR Principles as a frame for assessing the strengths and weaknesses of the speech and oral archives and scholarly practices in their field. Lastly, there have also been movements to extend FAIR and develop complementary principles such as the CARE Principles for Indigenous Data Governance, whose hashtag is `#BeFAIRandCARE`. The CARE Principles ask stewards and data ecosystems to take into account *collective benefit*, *authority to control*, *responsibility*, and *ethics* in their governance of Indigenous data (Global Indigenous Data Alliance, n.d.).

14 The authors of the original 2016 FAIR Principles quickly became aware of these scholarly communities' needs to translate the FAIR Principles into some type of metric to assess their projects, infrastructure, and practices and measure their progress towards achieving FAIRness:

> The Principles are aspirational, in that they do not strictly define how to achieve a state of 'FAIRness,' but rather they describe a continuum of features, attributes, and behaviors that will move a digital resource closer to that goal. This ambiguity has led to a wide range of interpretations of FAIRness, with some resources even claiming to already 'be FAIR'! The increasing number of such statements, the emergence of subjective and self-assessments of FAIRness, and the need of data and service providers, journals, funding agencies, and

> regulatory bodies to qualitatively or quantitatively evaluate such claims, led us to self-assemble and establish a FAIR Metrics group to pursue the goal of defining ways to measure FAIRness.
>
> (Wilkinson et al. 2018, 1)

15 After drafting its initial universal metrics, the FAIR Metrics working group quickly recognized that what is considered FAIR in one community may be quite different in another community (Wilkinson et al. 2018, 1). Thus, the working group accounted for this in allowing for a workflow for their universal metrics to be supplemented by domain- and community-specific metrics, including creating a template and GitHub repository for these communities to build upon and create their own metrics as well as the infrastructure for communities to contribute their metrics back to the open repository.

16 The creation of metrics by the working group has generated an opportunity for individual projects to now measure their FAIRness against a core set of universal metrics. For example, van Erp et al. were among the first to publish their use of the metrics to assess three open-source data catalogs: CKAN, Dataverse, and Invenio (van Erp et al. 2018). There is also now an opportunity for specific domains, such as the larger epigraphy community, to further develop and contribute back a set of community-specific FAIRness metrics.

3. IIP and FAIR Principles

17 When IIP originally started to incorporate linked open data URIs and to settle upon a robust archival format, the IIP team was not yet aware of the FAIR principles or FAIR metrics. We would now like to apply the FAIR metrics to the IIP archival format, which is based on practices and requirements that are particular to the epigraphic and the EpiDoc/TEI community. In the process, we can lay the groundwork for using metadata that are in the `<teiHeader>` to express the FAIR principles in a measurable way.

18 In 2016, when the Epigraphische Datenbank Heidelberg was threatened with having its funding withdrawn, Francisca Feraudi-Gruénais and Frank Grieshaber sent out a warning and call to action (Feraudi-Gruénais and Grieshaber 2016). This resulted in the formation of Epigraphy.info,[4] a community working to develop sustainable digital epigraphic practices in order to make sure that epigraphic research corpora are more broadly supported. Epigraphy.info recognizes the

significance of the FAIR principles as a way to ensure the survival of these research corpora. The principles help by extending traditional editorial conventions, canonical reference systems, and vocabularies and making them more broadly applicable, as well as by allowing for extensibility and growth. Encouraged by their embrace of FAIR, and bolstered by our efforts to generate an archival format for the IIP inscriptions, we will evaluate IIP encoding practices against the FAIR metrics, and extend our results to EpiDoc and TEI documents more generally.

4. Measuring FAIRness in IIP and EpiDoc

19 The template developed by Wilkinson et al. to document each FAIR metric is intended to be generally applicable across disciplines. Communities can work with the same template to describe the metrics in ways that are best applicable to their own data and practices. The template asks for the following nine pieces of information. Each field serves a different role: 1–2 and 9 identify, 3–5 provide rationale for measurement, and 6–8 describe how to measure each principle (Wilkinson et al. 2018, 2).

1. Metric identifier.
2. Metric name identifier.
3. To which principle does it apply?
4. What is being measured?
5. Why should we measure it?
6. What must be provided?
7. How do we measure it?
8. What is a valid result?
9. For which digital resource(s) is this relevant?

20 We will evaluate each metric to learn how well it measures FAIRness in IIP documents, and by extension, in EpiDoc and TEI. We assume at the outset that the IIP documents are valid with respect to the EpiDoc schema, adhere to the EpiDoc encoding guidelines, and use a `<teiHeader>` to encode metadata as recommended by the TEI Guidelines. The discussion below is based on these best practices and on the work we did to create archival versions of the working IIP documents; to apply the same metrics to other EpiDoc and TEI documents, they should also be valid according to the TEI schema and provide similar rich metadata.

Findable

21 The first five metrics evaluate the findability of resources. They focus on metadata and identifiers. An important and unique characteristic of TEI and EpiDoc files that affects these metrics is that data and metadata are usually part of the same document, and often in the same file. This removes the need to relate metadata records to data files, but also means that metadata alone are less visible.

FM-F1A: Identifier uniqueness

[https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_F1A.pdf]

22 This metric measures whether a resource's identifier is unique and described by a known identifier scheme. The IIP project partially meets this metric because it has a DOI that points to the project website and stands in for all the parts of the project. It would be more informative if there were an additional DOI either for the set of encoded archival source files, or for each individual document, as is the case for many scientific publications. The I.Sicily project,[5] which is collecting the complete corpus of inscriptions from ancient Sicily, mints a Zenodo DOI for each inscription in the corpus. As IIP inscriptions each have an identifier in the form `\w{4}\d{4}\w?` that is unique within the project, DOI suffixes would ideally incorporate this information together with a reference to the IIP corpus.[6] Using such an identifier scheme, an inscription could be referenced individually, outside the context of IIP, and at the same time be identifiable as belonging to the IIP corpus. The accuracy of a document-level DOI could also be checked against the IIP ID, which is encoded as an `<idno>` in the `<publicationStmt>`. Currently, the project-level DOI is referenced in the `<licence>` element of the document metadata.

23 We recommend assigning a DOI to each document, and encoding it using a second `<idno>` with the attribute `@type="DOI"`.

24 As the IIP documents are stored in an institutional repository, each file also has a persistent identifier (PID) provided by the repository in the form of a URI. The PID is not currently referenced in the IIP header metadata, but like the project ID, it can be encoded with `<idno type="BDR">`. Some projects have developed workflows that incorporate a repository-generated DOI back into the file, as the file is being deposited into a repository like Zenodo (Prag 2020; Wagner, n.d.). There may be other types of files beyond the epigraphic documents that should have unique identifiers: for example, local authority lists if they are external to the documents, or XSL scripts that can be used to process the files.

FM-F1B: Identifier persistence

[https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_F1B.pdf]

25 This metric addresses the universal problem of identifier persistence and how changing a URI might affect the findability of a resource. The FAIR metrics test for a policy that sets out how changes in identifier should be handled. If, as described above, there is a DOI for the data set as a whole or each of the individual documents in the corpus, then the DOI system itself provides the most secure form of persistence. "The system provides a means to continue interoperability through exchange of meaningful information about identified entities through at minimum persistence of the DOI name and a description of the referent" (Paskin 2009, 1591). In the case of IIP, this information is available at https://search.datacite.org/works/10.26300/pz1d-st89.

FM-F2: Machine-readability of metadata

[https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_F2.pdf]

26 This metric tests whether machine-readable metadata exist for the document in a format that will increase its findability. The metric is written assuming that the metadata record that is being tested is part of a metadata registry or can be discovered by a web search engine independently from the rest of the document. The FAIR principle for this metric is met by TEI (and therefore EpiDoc and IIP) documents, as they contain predictable, machine-readable metadata in the `<teiHeader>`. The namespace declaration identifies TEI as the data format. Within the TEI header, the `<fileDesc>` provides machine-actionable information about the digital object, the `<sourceDesc>` contains information about the object being encoded, the `<publicationStmt>` provides responsibility and licensing information, and the `<encodingDesc>` holds information about the project and its encoding practices.

27 This metric exposes an important characteristic of TEI documents which is demonstrated by the IIP archival format. Document metadata is closely tied to document data, often in a single file. It is machine-readable, but not necessarily available to search engines and indexes. And there is no central TEI-aware registry which would allow it to be easily located or natively indexed. FAIR discussions tend to draw on data from the life sciences, where there are more metadata registries and where metadata is created and resides separately from research data. IIP inscriptions are available to view on the web, but their source XML is less easily findable.

28 For TEI and EpiDoc, this metric could be better aligned with community practice if it were used to test for the existence of structured information in the `<teiHeader>` and for a namespace declaration. This information could also be exposed in a more general format in a registry or on the web. An EpiDoc-aware disciplinary registry can ingest TEI metadata directly and expose it to the web. The role and need for a discipline-specific registry will come up again in the following discussion.

FM-F3: Metadata clearly and explicitly include the identifier of the data they describe [https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_F3.pdf]

29 IIP, following the EpiDoc and TEI Guidelines, encodes the project identifier for a document in the `<idno>` element in the `<teiHeader>`, which seems to fulfill this metric. The FAIR metrics specify that the identifier take the form of a GUID[7] so the unique string can be searched in a document or on the web, and can be used to pair data and metadata files. For epigraphic corpora which often have well-developed identifier schemes, it is more relevant to modify the metric and recommend that each document have an ID that is unique within the project and to encode it in the `<idno>` element with an appropriate `@type` attribute. A more robust scheme that includes a project ID could be developed and documented by the larger community of digital epigraphy projects. It is interesting to note that the GUID, which is for all practical purposes unique, together with a web search engine, can play the same role as a registry.

FM-F4: (Meta)data are registered or indexed in a searchable resource [https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_F4.pdf]

30 This metric tests whether it is possible to find a document on the web using a search engine. In order for this to work, a metadata record, or—in the case of IIP or other EpiDoc documents—the file that contains the data and metadata, must be indexed by a search engine. The IIP corpus fails this test even though the documents contain metadata and identifiable IDs, because they are always accessed via a web application or from the institutional repository and therefore neither their unique IDs nor the PIDs assigned by the repository are indexed by web search engines. One solution is to allow the source files to be crawled, especially if they have unique identifiers, or to deposit them in a repository that exposes permanent identifiers and other metadata. Identifiers would also be findable if projects were encouraged to include identifiers in their web applications so

they could be crawled and indexed whether they are recognized as identifiers or not. To make IIP resources more findable, we will work with our institutional repository to expose more EpiDoc metadata and identifiers for indexing.

Accessibility

31 The second set of metrics focuses on accessibility, testing for the use of open protocols and permissions.

FM-A1.1: The protocol is open, free, and universally implementable [https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_A1.1.pdf]

FM-A1.2: The protocol allows for authentication and authorization where necessary [https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_A1.2.pdf]

32 The first two accessibility metrics can be treated together. IIP documents meet the first metric because they have no access restrictions and are accessible via HTTP, an open and well-documented protocol. Most epigraphic projects using EpiDoc provide their source data freely via HTTP. If a project's documents are not openly available, the second metric tests to see if the procedure for getting access is documented and if it works. Even in the case of openly available TEI and EpiDoc documents, this information should be specified in the `<availability>` and `<licence>` elements in the `<publicationStmt>`. In order for this information to be available, however, the document's metadata need to be exposed, or available in an index, catalog, or registry.

FM-A2: Metadata should be accessible even when the data are no longer available [https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_A2.pdf]

33 This metric is one that most corpora encoded using EpiDoc and TEI cannot easily meet. Currently, they maintain very rich technical and descriptive metadata in the `<teiHeader>`, which is part of the document, but there is no default, independent location that holds only metadata. If a corpus or document has a DOI, which in the case of IIP points to the project website, the DOI registry contains minimal, but persistent, metadata about the resource. This may prove that the document existed at one time and provide a hint about the contents, but is not otherwise very informative. This metric might be recast to make it more applicable to accessibility and persistence for TEI documents. For example, if one source of an inscription becomes unavailable, the conjoined TEI/EpiDoc metadata and data could continue to live by deposit in multiple places (LOCKSS).

Interoperability

34 The interoperability and reuse principles and associated metrics try to determine if, once retrieved, the data and metadata formats of a document are documented so they can be understood, and if the specifications on which they draw are themselves FAIR. Overall, IIP and other EpiDoc and TEI documents are fairly well positioned. The area that needs the most attention is not so much the encoded documents and corpora but rather the machine-verifiable FAIRness of the guidelines, schemas, and authorities on which they draw.

FM-I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

[https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_I1.pdf]

35 TEI and EpiDoc are XML schemas, so by extension, IIP files are expressed in XML. XML is an open standard, with a formal definition and an IANA (Internet Assigned Numbers Authority) media type. In addition, the TEI and EpiDoc schemas are written in ODD (One Document Does it all), a formal specification which is itself a TEI XML document that serves to link schema and documentation together (TEI Consortium 2020). There are two ways to point to the ODD in a TEI file: in the `<encodingDesc>` or in the `<?xml-model?>` processing instruction. The former is preferable because its presence can be ascertained using an XPath expression. IIP, as an EpiDoc project, should refer to the EpiDoc ODD in header metadata, but it currently does not include this information. Also, because the project strives to validate against the current EpiDoc schema in its day-to-day work, no EpiDoc version number is specified. In order to fulfill this metric, the IIP archival format should include this information in the `<encodingDesc>`.

FM-I2: (Meta)data use vocabularies that follow the FAIR principles

[https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_I2.pdf]

FM-I3: (Meta)data include qualified references to other (meta)data

[https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_I3.pdf]

36 The second and third interoperability metrics test for the use of well-defined and documented vocabularies, linked in semantically informative ways. They also recommend that linked data should point as often as possible to authorities that are external to the document or project. The IIP project links to public, documented, and stable authorities wherever possible. Geographic locations are linked to the Pleiades geographical gazetteer (Bagnall et al. 2006–), dates are linked to the PeriodO gazetteer of historical periods (Rabinowitz, Shaw, et al., n.d.), and object types are

disambiguated using links to the Getty Art and Architecture Thesaurus (Getty Research Institute, n.d.). Overall, IIP fulfills the criteria for interoperable and external vocabularies. However, these links are not explicitly qualified as specified in FM-I3. For example, the relationships are not expressed in RDF (Resource Description Framework), and so can only implicitly represent an identity relationship. Future enhancements to IIP metadata should take this into account. We will also be guided by the ongoing discussions around incorporating RDFa into TEI data and metadata. A positive sign is that the project was able to use metadata present in the TEI header to generate RDF triples in order to incorporate IIP into the Pelagios Network.[8] Epigraphy.info has also published an official draft of "Modeling Epigraphy with an Ontology" which outlines how to use existing ontologies such as CIDOC-CRM, Nomisma, and CRMtex[9] to describe inscribed objects (Bodard et al. 2021).

Reuse

37 Projects and documents that are encoded in TEI are generally well set up for both interoperability and reuse, as these were two goals of the TEI from its very beginnings. The `<teiHeader>` allows for including a license and providing provenance information both for source material, in the `<sourceDesc>`, and for the electronic document, in `<fileDesc>`. The existence of these elements can be tested using XPath expressions.

FM-R1.1: (Meta)data are released with a clear and accessible data usage license [https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_R1.1.pdf]

38 This test is fulfilled together with FM-A1.2, and can be tested by an XPath statement that resolves to the `<licence>` element in the `<publicationStmt>`. IIP documents use a CC BY-NC 4.0 license.

FM-R1.2: (Meta)data are associated with detailed provenance [https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_R1.2.pdf]

39 In the IIP archival format, both data and metadata have detailed, machine-readable provenance information. Authorship or responsibility for the data is available in the `<publicationStmt>`; proper citation information is provided in the `<availability>` element in the `<publicationStmt>`; and information about how and why the data or document was produced is encoded in the `<projectDesc>` and the `<encodingDesc>`, all of which are prescribed by the EpiDoc and based on TEI. In addition, the "how" question is partially answered in XML by the reference in the namespace declaration and in TEI by the `<schemaRef>` element. Currently the TEI

namespace declaration does not point to any further information about TEI. This metric would be even more adequately met if the link to the TEI namespace declaration also yielded pathways to viewing the TEI Guidelines and schema.

FM-R1.3: (Meta)data meet domain-relevant community standards [https://github.com/FAIRMetrics/Metrics/blob/master/Distributions/FM_R1.3.pdf]

40 IIP meets this last metric in principle through its adoption of EpiDoc, which has been developed by the epigraphic community for digital epigraphic corpora. However, the FAIR metrics look for a form of machine-verifiable certification that a document or dataset meets some minimum requirements with respect to metadata and formats. This is not only not available in the TEI context, it has also been rejected as a concept, as there is no governing body which might provide the certification, and in fact the Guidelines are, as indicated in their name, not a standard. However, if the TEI consortium and its members recommend how to fulfill the FAIR principles by using the `<teiHeader>` as discussed for each metric above, and provide XSLT and Schematron files for validation, the output of that file could indicate compliance. As always, it is the responsibility of the encoder and project to make sure that the metadata are accurate and detailed.

5. Conclusion and Future Work

41 We have described the modifications necessary in order to generate a more sustainable, archival form of the IIP corpus, and then evaluated them against the FAIR metrics. More modifications to the IIP archival format are necessary, such as minting DOIs for each document and adding the BDR PIDs into the `<teiHeader>` metadata. We have shown that the TEI header can provide much of the information required by the FAIR principles in a predictable and machine-readable form. Specifically, the TEI Guidelines and schema indicate where and how to encode licensing information, metadata formats, documentation, and identifiers and their presence can be verified using XSLT, XPath, and Schematron. Overall, the affordances of the TEI, best practices of the EpiDoc community, and IIP archival format decisions have resulted in a set of documents that measure up to the requirements of FAIR metrics. Furthermore, the best practices adopted by the IIP project over the course of its development incorporated many FAIR behaviors with little extra effort.

42 However, some issues that are unique to TEI encoding, as well as future directions of FAIR, indicate areas where the TEI community could do more. Generally, these are formal components that have to do primarily with persistence of metadata external to the IIP document, identifiers, and documentation of formats and vocabularies. In this paper, we have raised these issues with respect to each individual metric. One way to make metadata findable in a discipline-agnostic way is to deposit it into known registries such as ICPSR or other FAIR repositories.[10] Using such registries could be regarded as reducing accessibility to the full TEI and EpiDoc metadata. Differences between the metadata schemas in these repositories and the TEI metadata mean that the registered metadata would be less complete than the TEI header itself. A discipline-specific approach would be to create new registries, or to recommend an existing registry or aggregator which could undertake to expose more of the relevant disciplinary metadata. The EAGLE project attempted to do this for digital epigraphy,[11] but was not able to develop a sustainable financial model. Another registry-like effort is represented by the Trismegistos project (Depauw and Gheldof 2014), which is aggregating documents, metadata, people, places, and bibliography from the ancient world and assigning identifiers to them. Trismegistos has adopted a subscription model to sustain its work. IDEA has created a Zenodo community for epigraphy that contains epigraphic corpora, among other resources.[12]

43 Some discipline-specific issues are more thought-provoking. Like the LAM community, the text encoding and epigraphic communities are working with data that do not look like the life or physical sciences data for which FAIR was developed. TEI documents, unlike research data created in the sciences or even the social sciences, which may be observational or machine-generated, embody a more tightly bound mixture of data and metadata. In the TEI universe, linguistic corpora are perhaps the datasets most similar to those of the other sciences. They differ from literary and historical texts in their size, uniformity, and computability. Literary or historical texts mingle data and metadata, at least from the point of view of the scholar. Is a bibliography data or metadata? Is it different when encoded in the `<teiHeader>` than when it is in a `<back>` element? What about the physical description of a book or other text-bearing object? And what about the encoding itself, which is an interpretive, analytical overlay onto a more undifferentiated text (which itself inherently embodies an interpretation—in the case of epigraphy, even the letters of the text may represent an editorial intervention). FAIR metrics focus on metadata as a surrogate for a document

or dataset. IIP documents incorporate metadata about format, encoding, and the data themselves in one self-documenting entity, which makes them more FAIR, in spirit, if not in letter. This interconnection complicates conformance to metrics that are concerned with extracting metadata and treating them as surrogates for the complete document.

44 IIP followed the EpiDoc and TEI recommendations for encoding metadata about provenance and reuse, as well as metadata about the epigraphic object itself. Other epigraphic corpora may choose to handle some of these features differently, or not to include them at all. In order to make it easy for projects to make their documents and collections FAIR, there is a need for more prescriptive information than is currently in the TEI Guidelines, especially to ensure interoperability and reusability. This can most effectively be provided by disciplinary customizations like EpiDoc, which is used by a community with common requirements, and in situations where a group like Epigraphy.info can recommend external authorities, or document identifiers for epigraphic texts.

45 Finally, in the process of evaluating IIP against the FAIR metrics, we realized that not all FAIR criteria are under the control of a project; some depend on external entities. IIP relies on XML as an ISO standard, on the FAIRness of TEI and external vocabularies, as well as on the best practices recommended by the EpiDoc Guidelines. In order for a project like IIP to meet the FAIR metrics, it is necessary to also evaluate EpiDoc and TEI against the same criteria. According to the intent of the FAIR Metrics, the community represented by epigaphy.info or EpiDoc users as a whole should redefine the criteria of the fourteen metrics to better suit digital epigraphic documents. And as the FAIR Metrics Working Group recognized, universal metrics are not equally applicable to all domains; they have to be supplemented by domain- and community-specific metrics. Future work can continue to test and develop metrics that work best for individual projects using EpiDoc and TEI encoding and disciplinary communities that make use of them. We also look to developments in the epigraphic community and the TEI Consortium with respect to FAIRness at a broader level.

APPENDIXES

Appendix 1. Appendix: Summary of Metrics and Tests

Here we list some possible machine-actionable tests for the FAIR Metrics discussed above.

Metric	Test
FM-F1A: Identifier uniqueness	Check for DOI: `/TEI/teiHeader/fileDesc/publicationStmt/idno[@type="DOI"]`
FM-F1B: Identifier persistence	DOI registry information: https://search.datacite.org/works/10.26300/pz1d-st89.
FM-F2: Machine-readability of metadata	Check for machine-readable metadata in the document: https://www.w3.org/XML/ `/TEI/teiHeader`
FM-F3: Metadata clearly and explicitly include the identifier of the data they describe	Test for existence of DOI (or other identifiers): `/TEI/teiHeader/fileDesc/publicationStmt/idno[@type="DOI"]`
FM-F4: (Meta)data are registered or indexed in a searchable resource	Not currently testable
FM-A1.1: The protocol is open, free, and universally implementable	`/TEI/teiHeader/fileDesc/publicationStmt/availability`
FM-A1.2: The protocol allows for authentication and authorization where necessary	`/TEI/teiHeader/fileDesc/publicationStmt/availability/licence`
FM-A2: Metadata should be accessible even when the data are no longer available	LOCKSS, DOI registry, disciplinary registry
FM-I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	`/TEI/teiHeader/encodingDesc/schemaRef`
FM-I2: (Meta)data use vocabularies that follow the FAIR principles	Relies on FAIRness of vocabularies used.

FM-I3: (Meta)data include qualified references to other (meta)data	Not met
FM-R1.1: (Meta)data are released with a clear and accessible data usage license	IIP documents use a CC BY-NC 4.0 license: `/TEI/teiHeader/fileDesc/publicationStmt/availability/licence`
FM-R1.2: (Meta)data are associated with detailed provenance	`/TEI/teiHeader/fileDesc/titleStmt` `<?xml version="1.0" encoding="UTF-8"?>` `xmlns="http://www.tei-c.org/ns/1.0"` `/TEI/teiHeader/encodingDesc/schemaRef`
FM-R1.3: (Meta)data meet domain-relevant community standards	Can conform to recommendation, but without formal certification.

BIBLIOGRAPHY

Bagnall, Roger, Richard J. A. Talbert, Sarah Bond, Jeffrey Becker, Tom Elliott, Sean Gillies, Lindsay Holman, Ryan Horne, et al. n.d. “Pleiades.” Chapel Hill, NC: Ancient World Mapping Center, University of North Carolina; London: Stoa Consortium; New York: Institute for the Study of the Ancient World, New York University. Accessed July 6, 2022, https://pleiades.stoa.org/.

Bodard, Gabriel, Hugh Cayless, Chiara Cenati, Alison Cooley, Tom Elliott, Silvia Evangelisti, Achille Felicetti, et al. 2021. “Modeling Epigraphy with an Ontology.” Working paper, Ontology Working Group, Epigraphy.org, version 0.1, March 26. doi:10.5281/zenodo.4639507.

Calamai, Silvia, and Francesca Frontini. 2018. “FAIR Data Principles and Their Application to Speech and Oral Archives.” *Journal of New Music Research* 47 (4): 339–54. doi:10.1080/09298215.2018.1473449.

Data Citation Synthesis Group. 2014. “Joint Declaration of Data Citation Principles.” Edited by Martone M.. San Diego CA: FORCE11. doi:10.25490/a97f-egyk.

Depauw, Mark, and Tom Gheldof. 2014. “Trismegistos: An Interdisciplinary Platform for Ancient World Texts and Related Information.” In *Theory and Practice of Digital Libraries—TPDL 2013 Selected Workshops*, edited by Łukaz Bolikowski, Vittore Casarosa, Paula Goodale, Nikos Houssos, Paolo Manghi, and Jochen Schirrwagen, 40–52. Communications in Computer and Information Science 416. Cham: Springer. doi:10.1007/978-3-319-08425-1_5.

Elliott, Tom, Gabriel Bodard, Hugh Cayless, et al. 2021. "EpiDoc: Epigraphic Documents in TEI XML." Version 9.3, November 2021. http://epidoc.stoa.org.

Erp, Jarno A. A. van, Carolyn D. Langen, Anca Boon, and Kees van Bochove. 2018. "Testing the FAIR Metrics on Data Catalogs." Preprint, received and published September 4. PeerJ Preprints. doi:10.7287/peerj.preprints.27151v2.

European Commission, Directorate-General for Research and Innovation. 2016. *Realising the European Open Science Cloud: First Report and Recommendations of the Commission High Level Expert Group on the European Open Science Cloud.* Luxembourg: Publications Office of the European Union. doi:10.2777/940154.

———. 2018. *Turning FAIR into Reality: Final Report and Action Plan from the European Commission Expert Group on FAIR Data.* Luxembourg: Publications Office of the European Union. doi:10.2777/1524.

FAIRsFAIR. n.d. "The Project." Accessed June 8, 2020. https://www.fairsfair.eu/the-project.

Feraudi-Gruénais, Francisca, and Frank Grieshaber. 2016. "Digital Epigraphy am Scheideweg? / Digital Epigraphy at a Crossroads?" Nachnutzung und Nachnutzbarkeit der Forschung im Akademienprogramm Workshop der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste und der Union der deutschen Akademien der Wissenschaften AG "eHumanities." Düsseldorf: N.p, 2016. doi:10.11588/heidok.00022141.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. "Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics." *Scientific Data* 5, article 180205. doi:10.1038/sdata.2018.205.

Getty Research Institute. n.d. *Art & Architecture Thesaurus Online*. Accessed May 31, 2020. https://www.getty.edu/research/tools/vocabularies/aat/.

Global Indigenous Data Alliance. n.d. "CARE Principles for Indigenous Data Governance." Accessed June 9, 2020. https://www.gida-global.org/care.

Koster, Lukas, and Saskia Woutersen-Windhouwer. 2018. "FAIR Principles for Library, Archive and Museum Collections: A Proposal for Standards for Reusable Collections." *Code4Lib Journal* 40 (May). https://journal.code4lib.org/articles/13427.

Lamprecht, Anna-Lena, Leyla Garcia, Mateusz Kuzak, Carlos Martinez, Ricardo Arcila, Eva Martin Del Pico, Victoria Dominguez Del Angel, et al. 2020. "Towards FAIR Principles for Research Software." *Data Science* 3 (1): 37–59. doi:10.3233/DS-190026.

Leach, Paul J., Michael Mealling, and Rich Salz. 2005. "A Universally Unique IDentifier (UUID) URN Namespace." RFC 4122, Standards Track. N.p.: The Internet Society. https://www.ietf.org/rfc/rfc4122.txt.

Paskin, Norman. 2009. "Digital Object Identifier (DOI®) System." In *Encyclopedia of Library and Information Sciences*, 3rd ed., edited by Marcia J. Bates and Mary Niles Maack, 1586–92. Boca Raton: CRC Press. doi:10.1081/E-ELIS3.

Prag, Jonathan. 2020. "Publication in a Digital World." I.Sicily. December 23, 2020. https://isicily.org/2020/12/23/publication-in-a-digital-world/.

Rabinowitz, Adam, Ryan Shaw, et al. n.d. "PeriodO: A Gazetteer of Periods for Linking and Visualizing Data." Accessed May 31, 2020. http://perio.do/en/.

TEI Consortium. 2020. "Documentation Elements." Chap. 22 in *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.0.0. Last updated February 13. N.p.: TEI Consortium. https://www.tei-c.org/Vault/P5/4.0.0/doc/tei-p5-doc/en/html/TD.html.

Wagner, Andreas. n.d. "Tei2zenodo." Accessed June 8, 2020. https://gitlab.gwdg.de/rg-mpg-de/tei2zenodo.

Wilkinson, Mark D., Michel Dumontier, IJSbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3, article 160018. doi:10.1038/sdata.2016.18.

Wilkinson, Mark D., Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. 2018. "A Design Framework and Exemplar Metrics for FAIRness." *Scientific Data* 5, article 180118. doi:10.1038/sdata.2018.118.

NOTES

1 IIP website and search interface: Michael L. Satlow, "Inscriptions of Israel/Palestine," 2002– , accessed July 3, 2022, http://library.brown.edu/iip. Brown Digital Repository API documentation with examples: accessed July 3, 2022, https://library.brown.edu/iip/about/api/. IIP texts in GitHub: accessed July 3, 2022, https://github.com/Brown-University-Library/iip-texts/tree/master/epidoc-files.

2 Accessed July 3, 2022, https://repository.library.brown.edu/.

3 The first edition of deposited files with archival headers, last updated June 2019, can be viewed at https://repository.library.brown.edu/studio/collections/id_904/.

4 Accessed July 31, 2019, https://epigraphy.info/.

5 Accessed September, 2022 https://isicily.org/.

6 The digital journal *British Art Studies* (accessed July 3, 2022, http://britishartstudies.ac.uk/) assigns DOIs to each page in an article, and generates them in a form that functions as a human-readable citation as well as a machine-readable identifier. An example is Kelvin Chuah, "Instant Malaysia: Imagining a Nation at the Commonwealth Institute," *British Art Studies*, no. 13 (Sept. 2019): 4, https://doi.org/10.17658/issn.2058-5462/issue-14/kfein/p4.

7 Globally Unique Identifier or Universally Unique Identifier: see Leach, Mealling, and Salz (2005).

8 Pelagios network: accessed July 4, 2022, https://pelagios.org/. IIP RDF data: Brown University Library GitHub repository, accessed July 4, 2022, https://github.com/Brown-University-Library/iip-texts/blob/master/pelagios/iip-pelagios.ttl.

9 CIDOC (International Committee for Documentation) Conceptual Reference Model, accessed July 4, 2022, http://www.cidoc-crm.org/; Nomisma (knowledge organization system for numismatics), accessed July 4, 2022, http://nomisma.org/; CRMtex model for the study of ancient texts (an extension of CIDOC-CRM), accessed July 4, 2022, http://www.cidoc-crm.org/crmtex/home-8.

10 DataCite hosts a finder, developed in partnership with the American Geophysical Union, to locate repositories that are certified in the US as "FAIR Enabling" and in the EU as "FAIRsFAIR": accessed July 4, 2022, https://repositoryfinder.datacite.org/.

11 EAGLE, The Europeana network of Ancient Greek and Latin Epigraphy, accessed July 4, 2022, https://www.eagle-network.eu/.

12 International Digital Epigraphy Association on Zenodo, accessed July 4, 2022, https://zenodo.org/communities/eagle-idea.

AUTHORS

ANDREW CREAMER

Andrew Creamer is the scientific data management librarian in the Brown University Library. He assists faculty and student researchers with documenting their methods, describing their data, and providing the public and other researchers with long-term access to their research products via deposit in online repositories, including the Brown Digital Repository (BDR). Andrew also currently serves as an informationist for Brown's Superfund Center Community Engagement Core on a NIEHS-funded project focused on making the Center's research products more FAIR and enhancing their discovery, access, reuse, and citation.

https://orcid.org/0000-0002-5286

GAIA LEMBI

Gaia Lembi was a visiting assistant professor of Judaic Studies at Brown University and project manager for the Inscriptions of Israel/Palestine from 1995 to 2020.

https://orcid.org/0000-0001-8962

ELLI MYLONAS

Elli Mylonas works in the Center for Digital Scholarship in the Brown University Library. She is the technical manager of the Inscriptions of Israel/Palestine and is also part of the US Epigraphy project. Her experience with SGML, XML, and TEI is longstanding; she is currently a member of the EpiDoc maintenance group and has served several terms on the TEI Council. Mylonas was the managing editor of the Perseus Project when it first began, and has worked on many digital humanities and digital scholarship projects since then. Her background is in classics.

https://orcid.org/0000-0002-0215

MICHAEL SATLOW

Michael Satlow is professor of Judaic studies and religious studies at Brown University. He is the principal investigator of the Inscriptions of Israel/Palestine Project. He specializes in the history of Jews and Judaism in antiquity but also writes and teaches more broadly, including in digital humanities. Among other grants and fellowships, Satlow has been awarded a Humanities Open Book grant from the Mellon Foundation to digitize fifty volumes from the Brown Judaic Studies series and make them open access.

https://orcid.org/0000-0001-7692