Archiving a TEI project FAIRly

Andrew Creamer, Gail Lembi, Elli Mylonas, Michael Satlow

Brown University

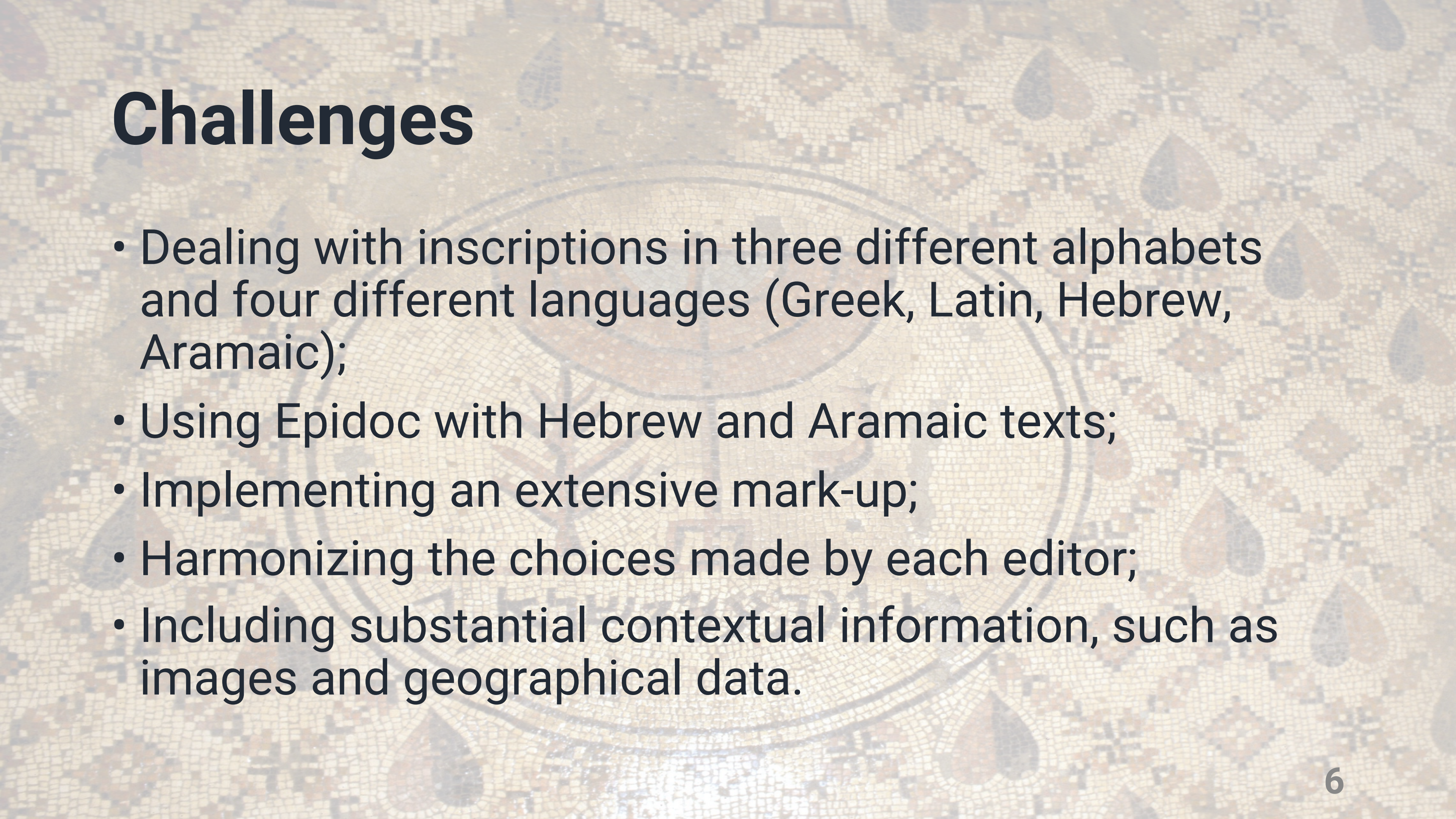
https://library.brown.edu/iip/index/

Inscriptions of Israel/Palestine: The Project

The Inscriptions of Israel/Palestine project (IIP) aims to build an internet accessible *corpus* of the inscriptions found in Israel/Palestine that date roughly between the VI century BCE and the VII century CE. These inscriptions are an invaluable resource for historical as well as linguistic investigation, since they provide information that is frequently not available through the extant literary sources: for instance, they reflect a broader social spectrum, convey religious views that have not been censored by a later normative tradition, and enhance our knowledge of diachronic developments and changes in the languages attested in the region.

Objectives

- To provide an exhaustive collection, that can be used by a variety of disciplines as well as by scholars and non scholars alike;
- To create an easily accessible platform which allows for extensive textual analysis;
- To allow for these data to be integrated with other contextual information;
- To link the corpus to other on line resources (e.g. Pelagios, Trismegistos, Pleiades, PeriodO).



Challenges

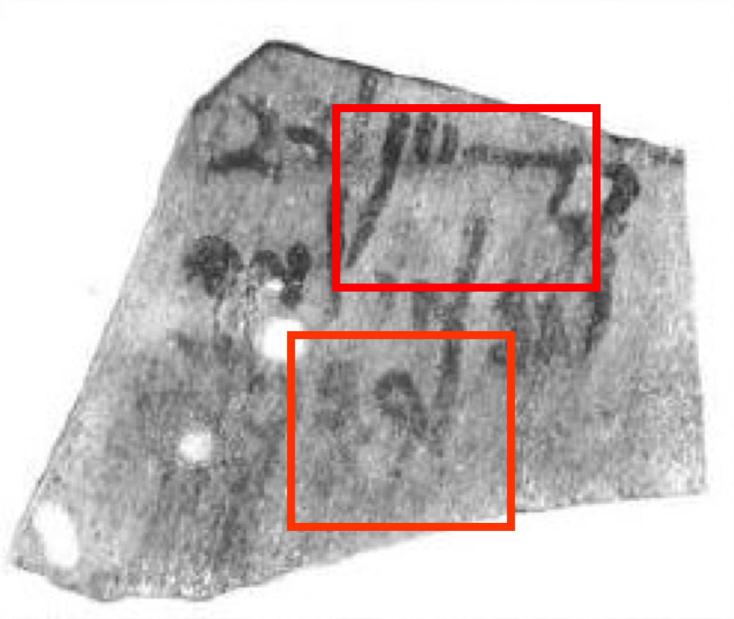
- Dealing with inscriptions in three different alphabets and four different languages (Greek, Latin, Hebrew, Aramaic);
- Using Epidoc with Hebrew and Aramaic texts;
- Implementing an extensive mark-up;
- Harmonizing the choices made by each editor;
- Including substantial contextual information, such as images and geographical data.

An example: An Ostracon from Masada

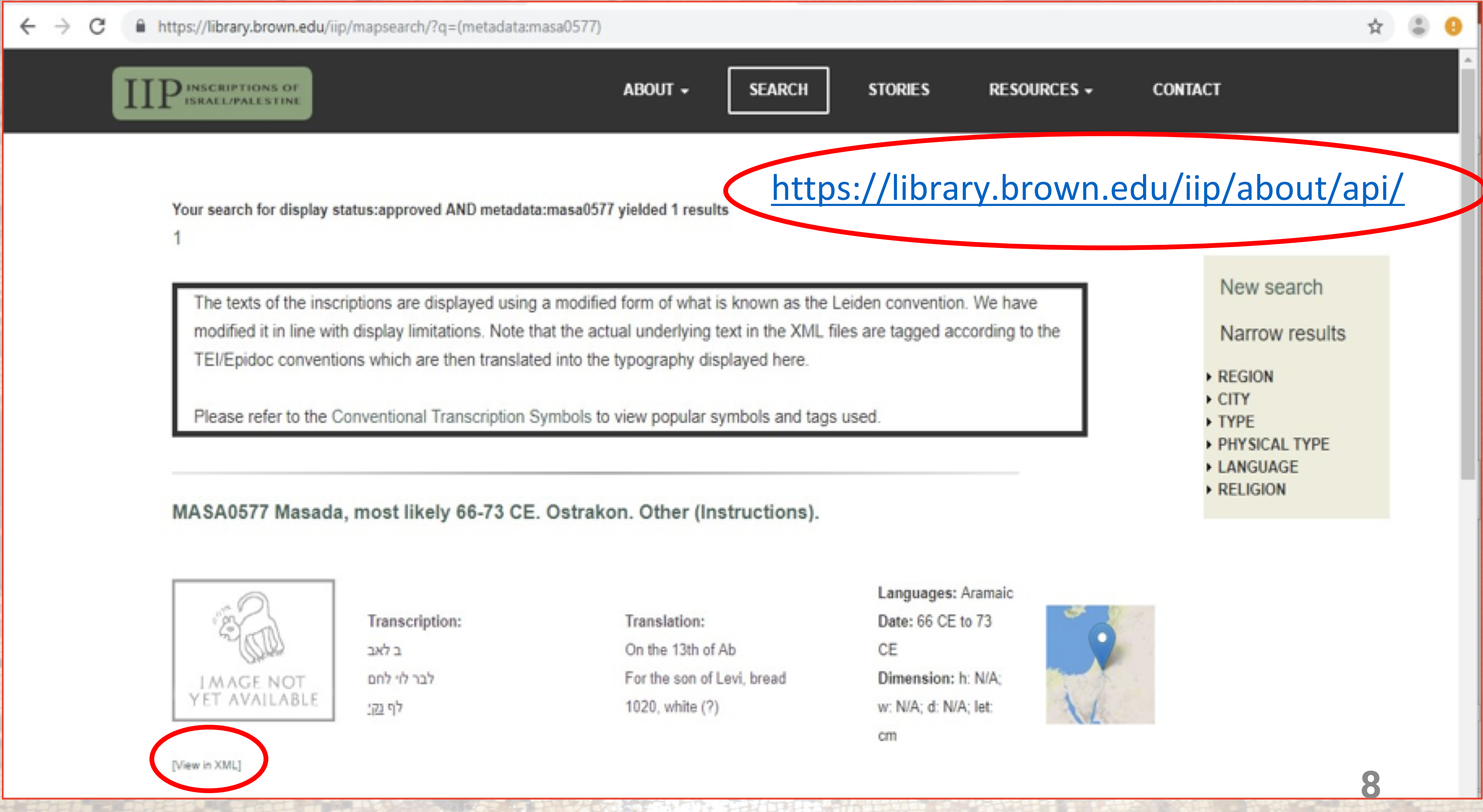
Example of the oddities faced while marking up such texts: the task is further complicated by the usage of paleo-Hebrew letters and numerals.

```
            <div type="edition"
subtype="transcription" ana="b1">
               <p xml:lang="heb">ב
<g ref="PHOENICIAN-NUMBER-TEN"/><g
ref="PHOENICIAN-NUMBER-THREE"
                   /> לאב<lb/> לבר
לוי לחם<lb/><g ref="PHOENICIAN-
NUMBER-1000">לף</num><g
ref="PHOENICIAN-NUMBER-TWENTY"/>

<unclear>נקי</unclear></p></div>
```



“On the 13 of Av / For the son of Levi, bread / 1020, white (?)”



https://library.brown.edu/iip/about/api/

Your search for display status:approved AND metadata:masa0577 yielded 1 results

1

The texts of the inscriptions are displayed using a modified form of what is known as the Leiden convention. We have modified it in line with display limitations. Note that the actual underlying text in the XML files are tagged according to the TEI/Epidoc conventions which are then translated into the typography displayed here.

Please refer to the Conventional Transcription Symbols to view popular symbols and tags used.

MASA0577 Masada, most likely 66-73 CE. Ostrakon. Other (Instructions).

Transcription:
ב לאב
לבר לוי לחם
לף נק:

Translation:
On the 13th of Ab
For the son of Levi, bread
1020, white (?)

Languages: Aramaic
Date: 66 CE to 73 CE
Dimension: h: N/A; w: N/A; d: N/A; let: cm

[View in XML]

Working with the Texts: NLP

Possible outputs:

- Lemmatization:
  a. Concordances;
  b. Marking-up names and linking them to onomastica;
  c. Linking to dictionaries (e.g. Ma’agarim);
- Finding a way to deal with misspelled/misused words;
- Proposing reconstructions alternative to the editors’ ones.

Challenges:

- Scarcity of relevant tools for Semitic languages, especially Aramaic;
- Applying syntactic parsing to texts that span over many centuries.

Epidat: Terms Contained in <p>

```
<objectDesc>
   <supportDesc>
      <support>
        <p>
          <material>Stein</material>
          <objectType 
ref="http://vocab.getty.edu/page/aat/300005909">Grabmal
</objectType>
        <dimensions>
          <dim>Laut Gildemeister ist "der von der 
Inschrift eingenommene Theil [...] zwei Fuß breit und 
drei hoch".</dim>
        </dimensions>
      ...
```

http://www.steinheim-institut.de/cgi-bin/epidat?lang=en&release=beta

```
<supportDesc>
  <support>
   <objectType xml:lang="ru"    ref="monument-
   search.xml#mon1">Плита.</objectType>
   <objectType xml:lang="en">Panel.</objectType>
   <material xml:lang="ru" ref="material-
   search.xml#m1">Мрамор белый, сероватый.</material>
   <material xml:lang="en">White-grayish
   marble.</material>
   <p xml:lang="ru">Лицевой фас отшлифован, оборотный
   обработан грубыми сколами, остальные оббиты.</p>
   <p xml:lang="en">…</p>
  </support>
 </supportDesc>
```

http://iospe.kcl.ac.uk/index.html

IIP: terms in @ana followed by <p>

```
<objectDesc ana="#amphora #handles">
   <supportDesc ana="#clay">
      <support>
         <p>Complete vertical part of a handle;
rounded profile of the top. The fabric has a
gritty texture and is light yellowish red. The
surface is very pale brown, with fine yellow
mica</p>
      </support>
   </supportDesc>
</objectDesc>
```

IIP template

```
<msDesc>
        <msIdentifier/>
        <msContents>
                <textLang mainLang="" otherLangs=""/>
                <msItem class="" ana="">
                        <p></p>
                </msItem>
        </msContents>
        <physDesc>
                 <objectDesc ana="">
                        <supportDesc ana="">
                                <support>
                                        <p/>
                                </support>
                        </supportDesc>
                </objectDesc>
 </msDesc>
```

Linking Textual Derived Information

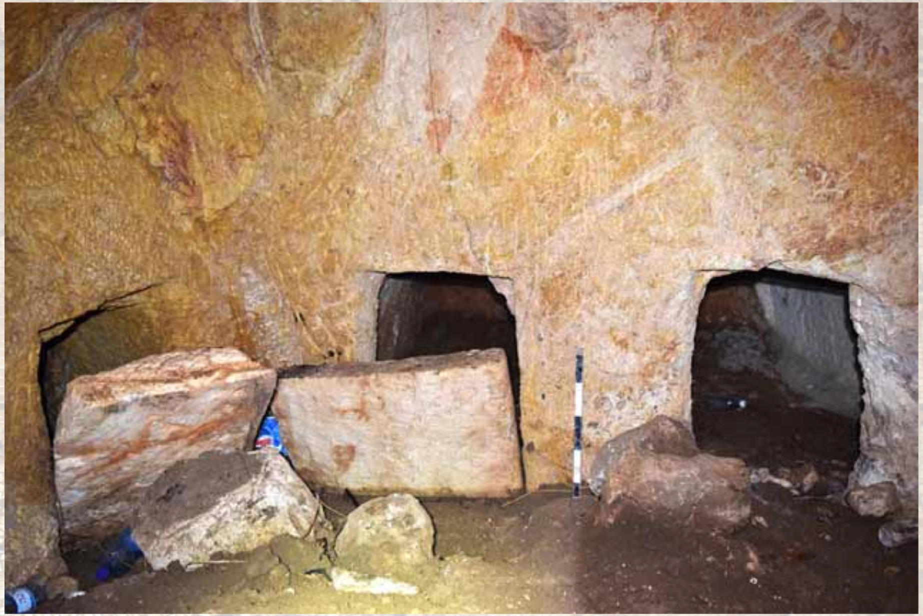
- Pleiades
  1. Mutually beneficial: we linked our material to the Gazetteer, while contributing a fair amount of new places to it;
  2. Geographical information could be analyzed through GIS to study a large variety of topics, from the diachronic distribution of the inscriptions, to the relationship between language and religion to name just a few.
- Getty Art and Architecture Thesaurus
  1. Standardization is still a *desideratum* for controlled vocabularies;
  2. Choosing Getty AAT over EAGLE vocabularies: the challenge of working with a tool that was not created with epigraphy in mind.

Kokh

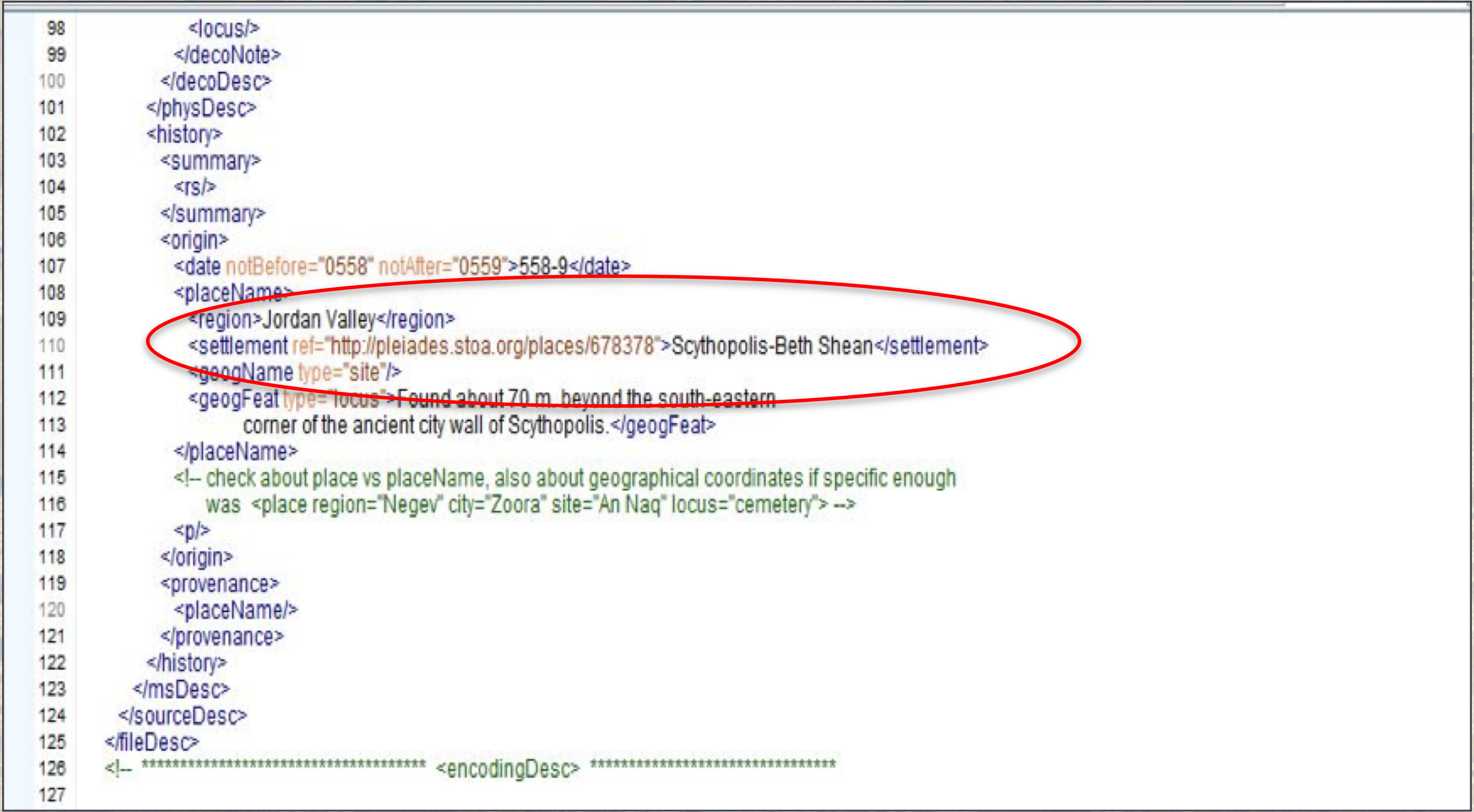
A kokh (plural: kokhim, Hebrew: כּוּךְ) is a type of tomb complex characterized by a series of long narrow shafts, in which the deceased were placed for burial, radiating from a central chamber. These tomb complexes were generally carved into a rock face, and were usually closed with a stone slab and had channels cut into the centre of the shaft to drain any water that seeped through the rock.

https://en.wikipedia.org/wiki/Rock-cut_tomb#Kokh

Not in Getty AAT
Not in EAGLE vocabularies

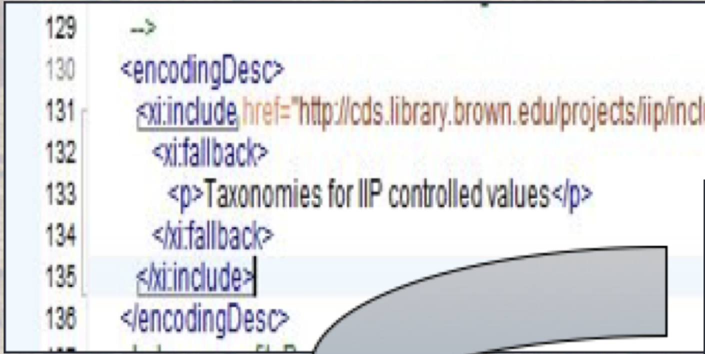


http://www.hadashot-esi.org.il/Report_Detail_Eng.aspx?id=25240

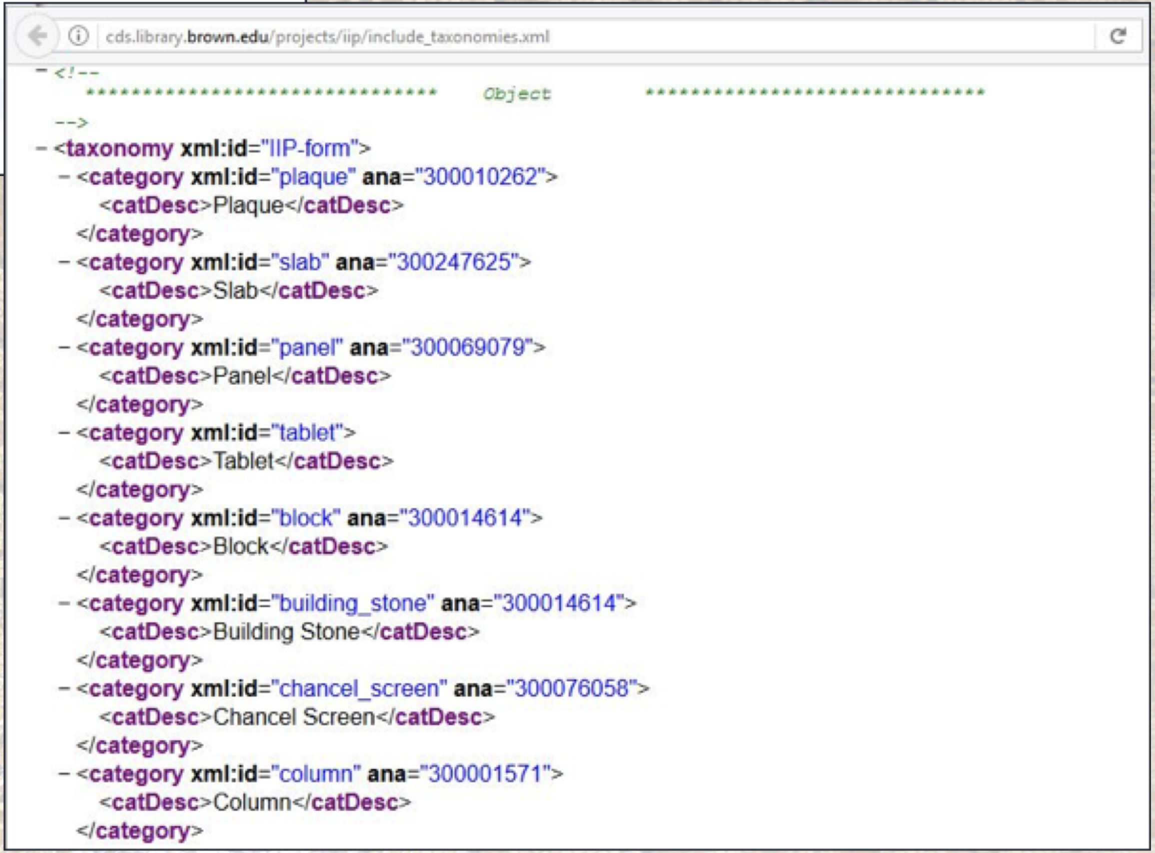


```
            <locus/>
          </decoNote>
        </decoDesc>
      </physDesc>
      <history>
        <summary>
          <rs/>
        </summary>
        <origin>
          <date notBefore="0558" notAfter="0559">558-9</date>
          <placeName>
            <region>Jordan Valley</region>
            <settlement ref="http://pleiades.stoa.org/places/678378">Scythopolis-Beth Shean</settlement>
            <geogName type="site"/>
            <geogFeat type="locus">Found about 70 m. beyond the south-eastern
                    corner of the ancient city wall of Scythopolis.</geogFeat>
          </placeName>
          <!-- check about place vs placeName, also about geographical coordinates if specific enough
              was  <place region="Negev" city="Zoora" site="An Naq" locus="cemetery"> -->
          <p/>
        </origin>
        <provenance>
          <placeName/>
        </provenance>
      </history>
    </msDesc>
  </sourceDesc>
</fileDesc>
<!--  ****************************************  <encodingDesc>  *********************************

```



```
  -->
 <encodingDesc>
   <xi:include href="http://cds.library.brown.edu/projects/iip/include_taxonomies.xml">
     <xi:fallback>
       <p>Taxonomies for IIP controlled values</p>
     </xi:fallback>
   </xi:include>
 </encodingDesc>
```



cds.library.brown.edu/projects/iip/include_taxonomies.xml

```
- <!--
      ********************************     Object        ********************************
  -->
- <taxonomy xml:id="IIP-form">
  - <category xml:id="plaque" ana="300010262">
      <catDesc>Plaque</catDesc>
    </category>
  - <category xml:id="slab" ana="300247625">
      <catDesc>Slab</catDesc>
    </category>
  - <category xml:id="panel" ana="300069079">
      <catDesc>Panel</catDesc>
    </category>
  - <category xml:id="tablet">
      <catDesc>Tablet</catDesc>
    </category>
  - <category xml:id="block" ana="300014614">
      <catDesc>Block</catDesc>
    </category>
  - <category xml:id="building_stone" ana="300014614">
      <catDesc>Building Stone</catDesc>
    </category>
  - <category xml:id="chancel_screen" ana="300076058">
      <catDesc>Chancel Screen</catDesc>
    </category>
  - <category xml:id="column" ana="300001571">
      <catDesc>Column</catDesc>
    </category>
```

As the screenshots show, each XML file refers to our external authority list for object types, which contains the corresponding Getty's ID.

Rights Statement

```
<publicationStmt>
   <authority>Brown University</authority>
   <idno type="IIP">beth0023</idno>
   <availability status="free">
     <licence>This work is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License. <ref
target= "http://creativecommons.org/licenses/by-nc/4.0/">
Distributed under a Creative Commons licence CC BY-NC 4.0</ref>
      </licence>
    </availability>
</publicationStmt>
```

Edition and Citation info

```
<editionStmt>
    <edition n="v1">First archival edition
        <date>2019-03-15</date>
    </edition>
</editionStmt>
```

[in <availability>]

```
<p>All reuse or distribution of this work must contain
somewhere a link to the DOI of the Inscriptions of
Israel/Palestine Project:
<ref>https://doi.org/10.26300/pz1d-st89</ref>
</p>
```

Internal authority lists in `//encodingDesc/classDecl`

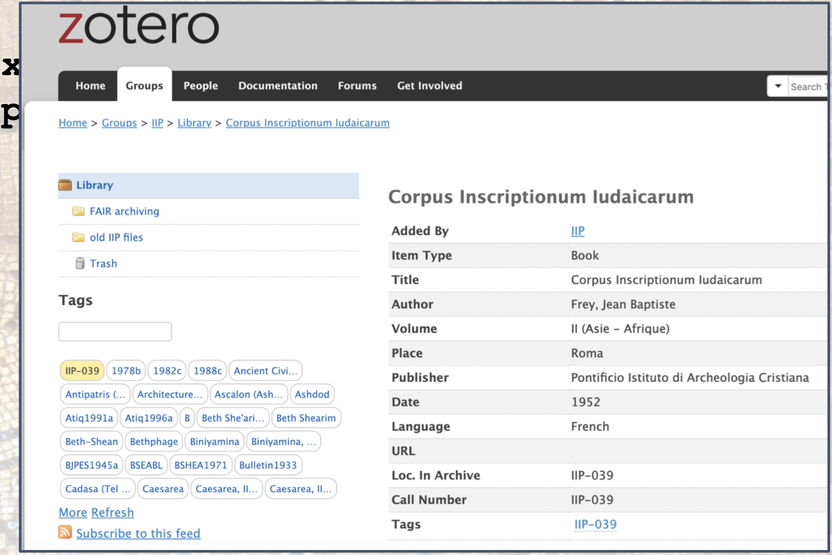
```
<encodingDesc>
     <xi:include
href="http://cds.library.brown.edu/projects/iip/include_taxonomies.xml">
        <xi:fallback>
         <p>ERROR: could not find taxonomies file, which should appear
             in this space.</p>
        </xi:fallback>
    </xi:include>
</encodingDesc>
```

Internal authority lists in `//encodingDesc/classDecl`

```
<taxonomy xml:id="IIP-form">
    <category xml:id="arch" ana="300000994">
        <catDesc>Arch</catDesc>
    </category></taxonomy>
 <taxonomy xml:id="IIP-genre">
     <category xml:id="funerary">
        <catDesc>Funerary</catDesc>
    </category></taxonomy>
  <taxonomy xml:id="IIP-preservation">
     <category xml:id="complete.intact">
        <catDesc>Complete and intact</catDesc>
    </category></taxonomy>
  <taxonomy xml:id="IIP-writing">
     <category xml:id="painted">
        <catDesc>Painted</catDesc>
     </category></taxonomy>
  <taxonomy xml:id="IIP-religion">
     <category xml:id="jewish">
        <catDesc>Jewish</catDesc>
    </category></taxonomy>
```

Bibliography in //back/div[@type=”bibliography”]

```
<div type="bibliography">
   <listBibl>
      <bibl xml:id="b2">
        <ptr type="biblItem" target="IIP-039.xml"/>
        <biblScope unit="page">8</biblScope>
      </bibl>
      <bibl xml:id="b3">
        <ptr type="biblItem" target="IIP-145.x
        <biblScope unit="page">35-36</biblScop
      </bibl>
   </listBibl>
</div>
```



Home > Groups > IIP > Library > Corpus Inscriptionum Iudaicarum

Library
- FAIR archiving
- old IIP files
- Trash

Tags

IIP-039 · 1978b · 1982c · 1988c · Ancient Civi... · Antipatris (... · Architecture... · Ascalon (Ash... · Ashdod · Atiq1991a · Atiq1996a · B · Beth She'ari... · Beth Shearim · Beth-Shean · Bethphage · Biniyamina · Biniyamina, ... · BJPES1945a · BSEABL · BSHEA1971 · Bulletin1933 · Cadasa (Tel ... · Caesarea · Caesarea, II... · Caesarea, II...

More Refresh

Subscribe to this feed

Corpus Inscriptionum Iudaicarum

Added By	IIP
Item Type	Book
Title	Corpus Inscriptionum Iudaicarum
Author	Frey, Jean Baptiste
Volume	II (Asie – Afrique)
Place	Roma
Publisher	Pontificio Istituto di Archeologia Cristiana
Date	1952
Language	French
URL	
Loc. In Archive	IIP-039
Call Number	IIP-039
Tags	IIP-039

Bibliography in //back/div[@type=”bibliography”]

```
<div type="bibliography">
   <biblStruct n="IIP-039" xml:id="b1">
      <monogr>
         <title level="m">Corpus Inscriptionum Iudaicarum</title>
         <idno type="callNumber">IIP-039</idno>
         <author><forename>Jean Baptiste</forename><surname>Frey</surname></author>
          <imprint>
          <pubPlace>Roma</pubPlace>
          <biblScope unit="volume">II (Asie - Afrique)</biblScope>
          <publisher>Pontificio Istituto di Archeologia Cristiana</publisher>
          <date>1952</date>
          </imprint>
       </monogr>
       <citedRange unit="insc">403</citedRange>
    </biblStruct>
    <biblStruct n="IIP-403" xml:id="b2">
        <analytic>
           <title level="a">Greek Inscriptions from Beth She'arim</title>
           <author><forename>Moshe</forename><surname>Schwabe</surname></author>
        </analytic>…</biblStruct></div>
```

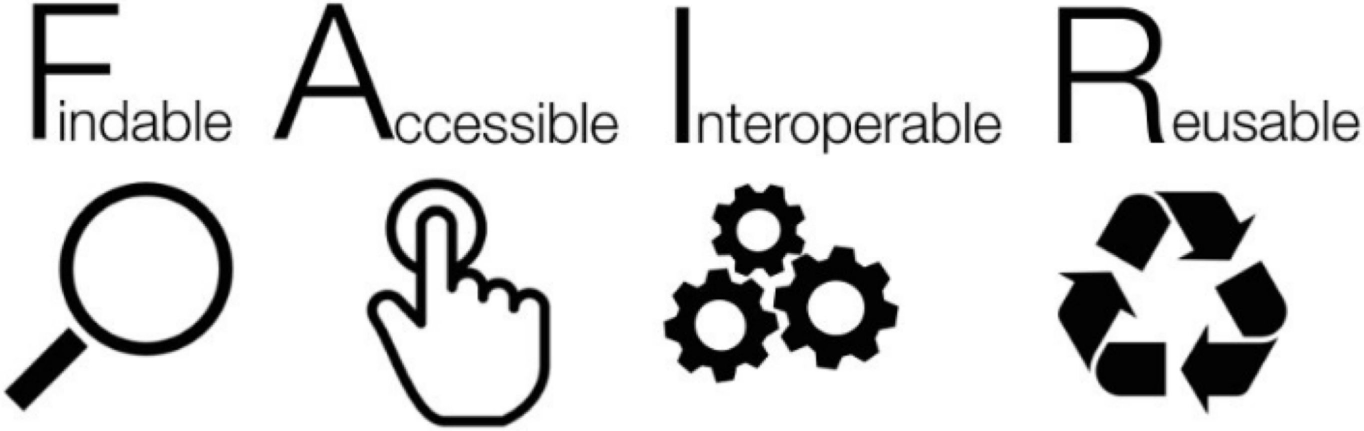
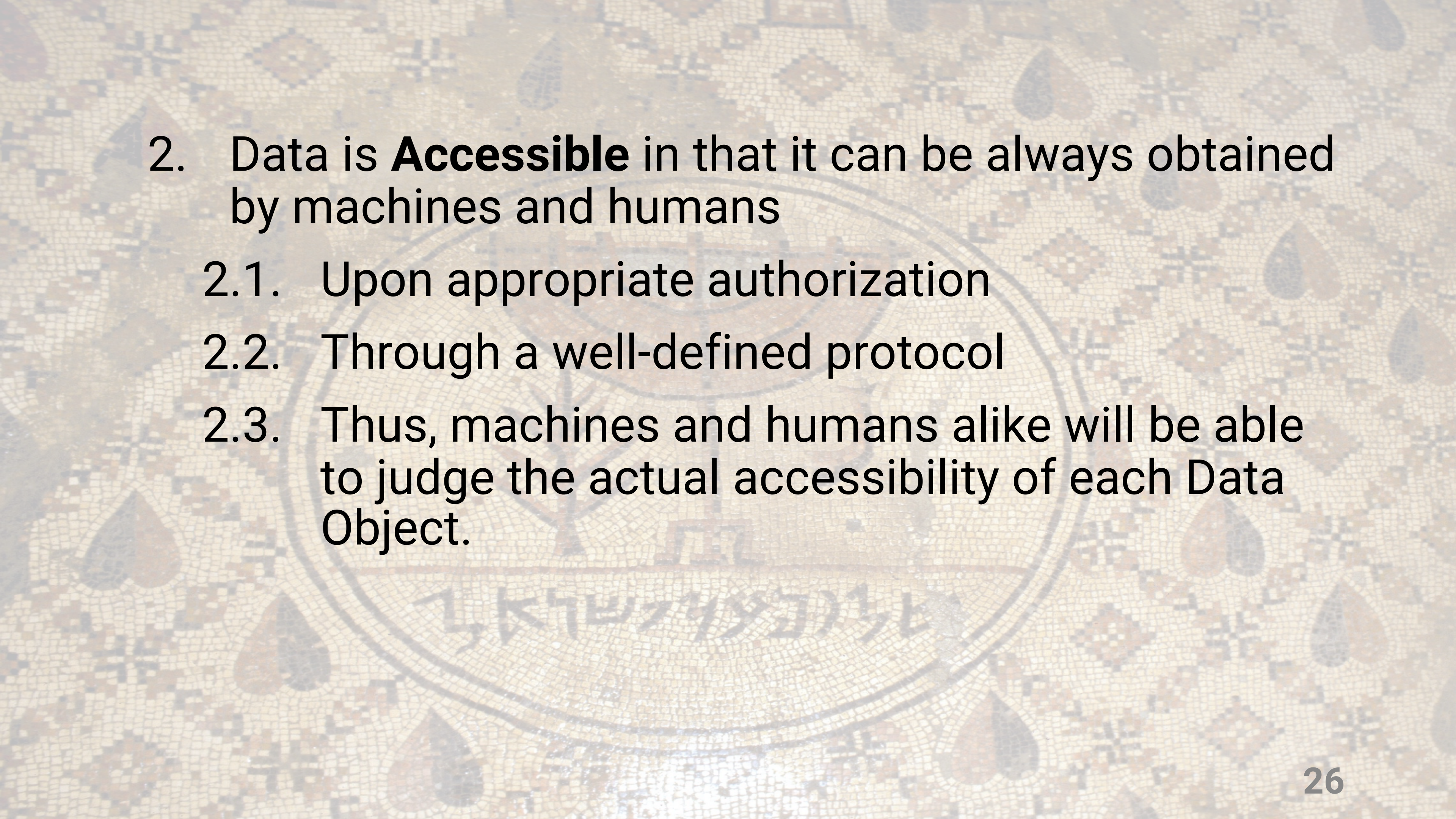
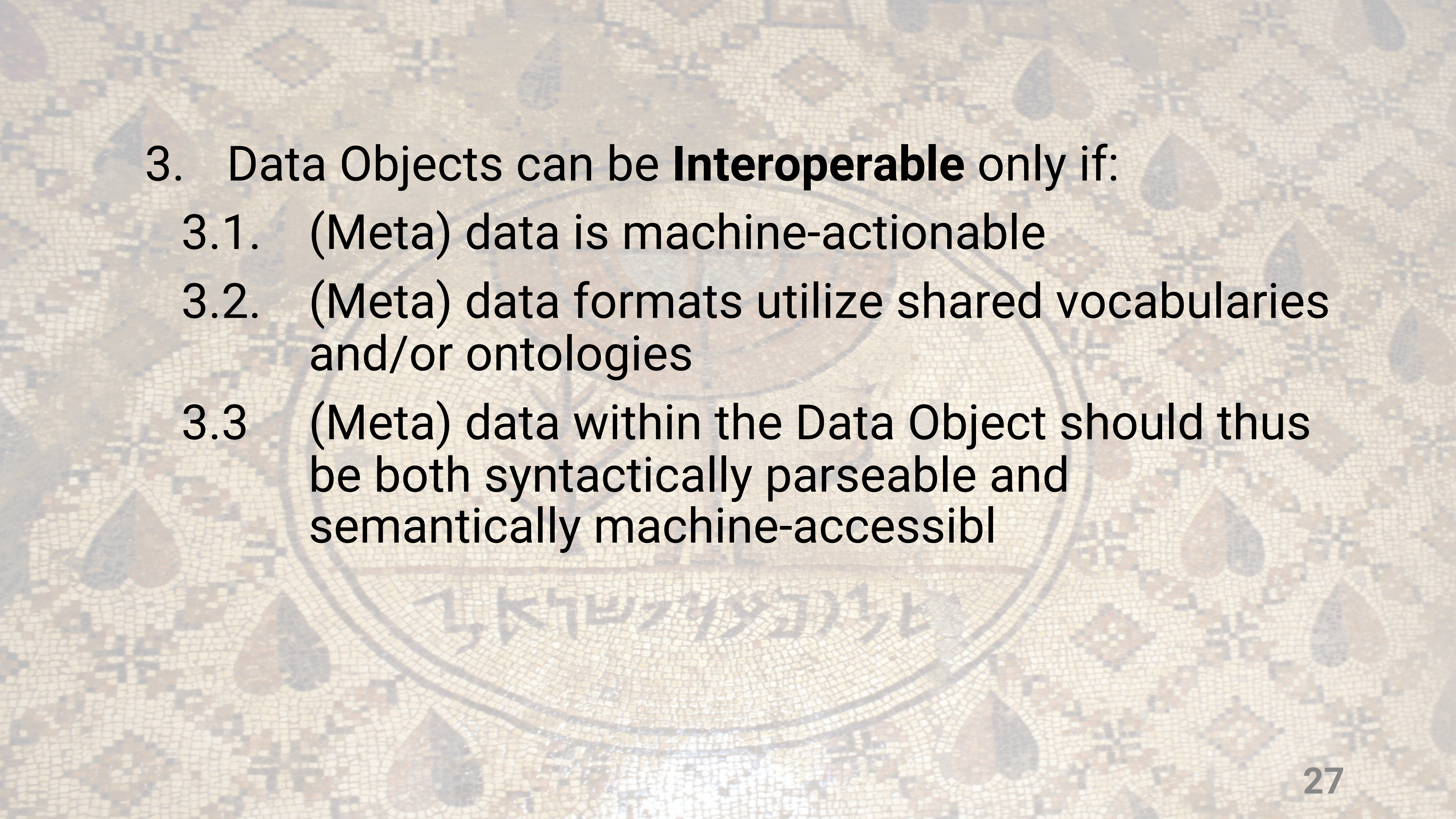


Image credit: Sungya Pundir CCBYSA4.0

1. To be **Findable** any Data Object should be uniquely and persistently identifiable
    1. The same Data Object should be re-findable at any point in time, thus Data Objects should be persistent, with emphasis on their metadata,
    2. A Data Object should minimally contain basic machine actionable metadata that allows it to be distinguished from other Data Objects
    3. Identifiers for any concept used in Data Objects should therefore be Unique and Persistent



2. Data is **Accessible** in that it can be always obtained by machines and humans
   2.1. Upon appropriate authorization
   2.2. Through a well-defined protocol
   2.3. Thus, machines and humans alike will be able to judge the actual accessibility of each Data Object.



3. Data Objects can be **Interoperable** only if:
   - 3.1. (Meta) data is machine-actionable
   - 3.2. (Meta) data formats utilize shared vocabularies and/or ontologies
   - 3.3 (Meta) data within the Data Object should thus be both syntactically parseable and semantically machine-accessibl

4. For Data Objects to be **Re-usable** additional criteria are:

4.1 Data Objects should be compliant with principles 1-3

4.2 (Meta) data should be sufficiently well-described and rich that it can be automatically (or with minimal human effort) linked or integrated, like-with-like, with other data sources

4.3 Published Data Objects should refer to their sources with rich enough metadata and provenance to enable proper citation

Example: Adapting FAIR principles for LAM

- Evaluate FAIR and other related guidelines
- Adapt to particular needs

The two main limitations are the lack of explicit attention for long term preservation of digital objects, besides their metadata, and the excessive interwovenness of “data” (or objects) and “metadata”

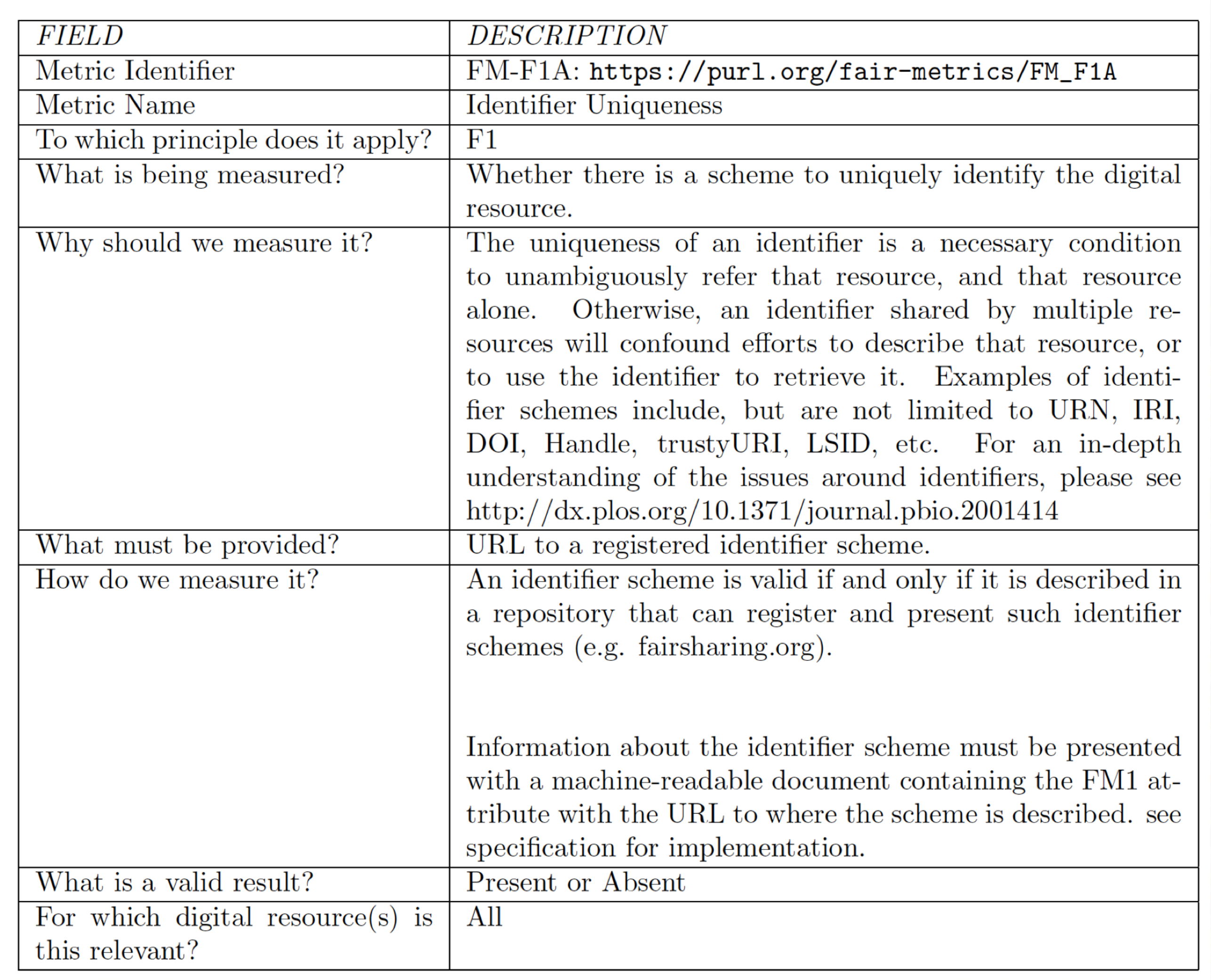
- Generate distinct FAIR principles for:
  - Objects
  - Metadata
  - Metadata Records

Koster, L., & Woutersen-Windhouwer, S. (2018). FAIR Principles for Library, Archive and Museum Collections: A proposal for standards for reusable collections. Code4Lib Journal, 40

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. A design framework and exemplar metrics for FAIRness doi: [10.1038/sdata.2018.118](https://doi.org/10.1038/sdata.2018.118)

14 Universal Exemplar FAIR Metrics



FIELD	*DESCRIPTION*
Metric Identifier	FM-F1A: `https://purl.org/fair-metrics/FM_F1A`
Metric Name	Identifier Uniqueness
To which principle does it apply?	F1
What is being measured?	Whether there is a scheme to uniquely identify the digital resource.
Why should we measure it?	The uniqueness of an identifier is a necessary condition to unambiguously refer that resource, and that resource alone. Otherwise, an identifier shared by multiple resources will confound efforts to describe that resource, or to use the identifier to retrieve it. Examples of identifier schemes include, but are not limited to URN, IRI, DOI, Handle, trustyURI, LSID, etc. For an in-depth understanding of the issues around identifiers, please see http://dx.plos.org/10.1371/journal.pbio.2001414
What must be provided?	URL to a registered identifier scheme.
How do we measure it?	An identifier scheme is valid if and only if it is described in a repository that can register and present such identifier schemes (e.g. fairsharing.org). Information about the identifier scheme must be presented with a machine-readable document containing the FM1 attribute with the URL to where the scheme is described. see specification for implementation.
What is a valid result?	Present or Absent
For which digital resource(s) is this relevant?	All

F1A: Identifier Uniqueness. F1B: Identifier Persistence. (F1: (Meta)data are assigned globally unique and persistent identifiers) Metadata Longevity. (A2: Metadata should be accessible even when the data is no longer available)

- IIP is identified by a DOI (Unique, Persistent)
- Archival IIP documents in the Brown Digital Repository have persistent identifiers. (Unique, Persistent)
- The DOI will persist even if the document(s) are no longer extant (Longevity)

Questions:

- What is the optimal granularity for the DOI or DOI+suffix?
- What else should have a DOI?
- What information should remain if the document disappears?
- What kind of policy should we have and how should we proceed if DOIs no longer exist?

Machine-readability of metadata (F2: Data are described with rich metadata)
Use a Knowledge Representation Language (I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation)

- Use of the <teiHeader> provides file, object metadata and data in machine readable form. IIP exposes some of them via its API
- IIP includes external references to local authority files within each document in the corpus.
- IIP should fully identify or refer to external references (Pleiades, Getty AAT).
- Data are richly described so that they can be interoperable and re-used.
- <teiHeader> has a mix of structured and unstructured information that could make it less machine actionable.
- No checksum
- TEI and Epidoc could be but are not FAIR themselves

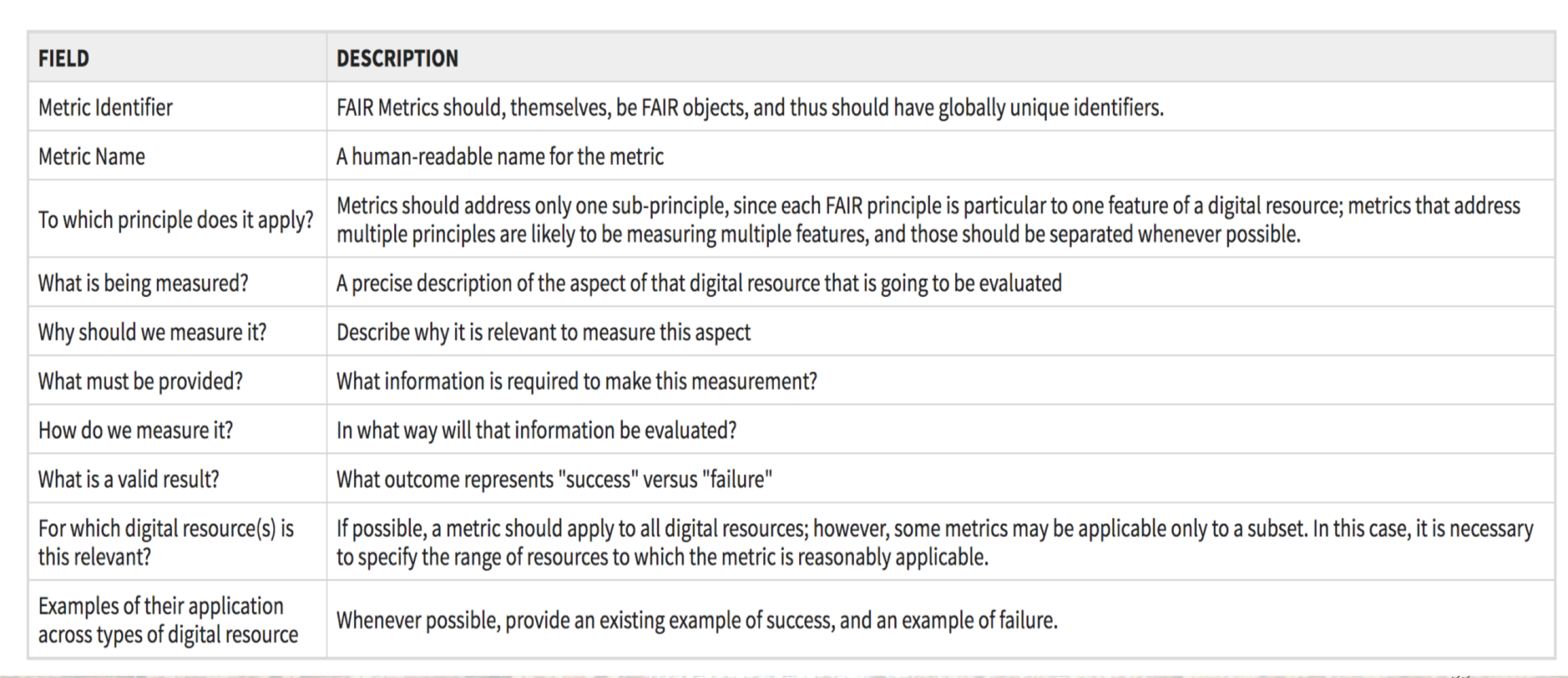
Indexed in a searchable resource (F4: (Meta)data are registered or indexed in a searchable resource)

Use FAIR Vocabularies (I2: (Meta)data use vocabularies that follow the FAIR principles)

Meets Community Standards (R1.3: (Meta)data meet domain-relevant community standards)

- IIP is indexed in the Brown Digital Repository
- Not indexed in Google because the documents are accessed as search results. Possible in Google Scholar?
- IIP uses shared community vocabularies and LOD, but it’s not clear that they are FAIR
- Overall, IIP relies on data structures and vocabularies developed by the epigraphic community where possible.

FAIR Accessor: A future direction for epigraphers?



FIELD	DESCRIPTION
Metric Identifier	FAIR Metrics should, themselves, be FAIR objects, and thus should have globally unique identifiers.
Metric Name	A human-readable name for the metric
To which principle does it apply?	Metrics should address only one sub-principle, since each FAIR principle is particular to one feature of a digital resource; metrics that address multiple principles are likely to be measuring multiple features, and those should be separated whenever possible.
What is being measured?	A precise description of the aspect of that digital resource that is going to be evaluated
Why should we measure it?	Describe why it is relevant to measure this aspect
What must be provided?	What information is required to make this measurement?
How do we measure it?	In what way will that information be evaluated?
What is a valid result?	What outcome represents "success" versus "failure"
For which digital resource(s) is this relevant?	If possible, a metric should apply to all digital resources; however, some metrics may be applicable only to a subset. In this case, it is necessary to specify the range of resources to which the metric is reasonably applicable.
Examples of their application across types of digital resource	Whenever possible, provide an existing example of success, and an example of failure.

Thank you!

gaia_lembi@brown.edu

elli_mylonas@brown.edu @elli_m