UNIVERSITÄT ZU LÜBECK

Information Systems

CS4130-KP06

Prof. Dr. Sylvia Melzer

SoSe2026

Data Models & Data Representation

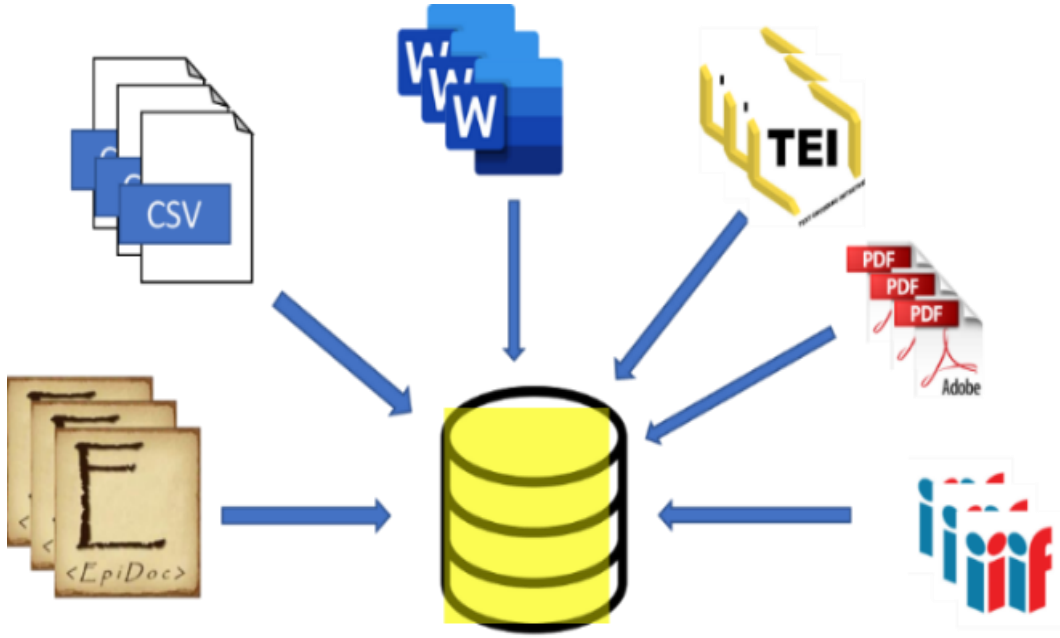
Information Systems

What You Will Learn

- What is a data model and why it matters
- To explain differences between unstructured and structured data
- To recognize TEI and EpiDoc as modeling approaches
- To interpret how representation formats encode meaning
- To understand the role of schemas such as Relax NG
- To transform structured text into tabular formats
- To reflect on underlying data models

The Nature of Research Data

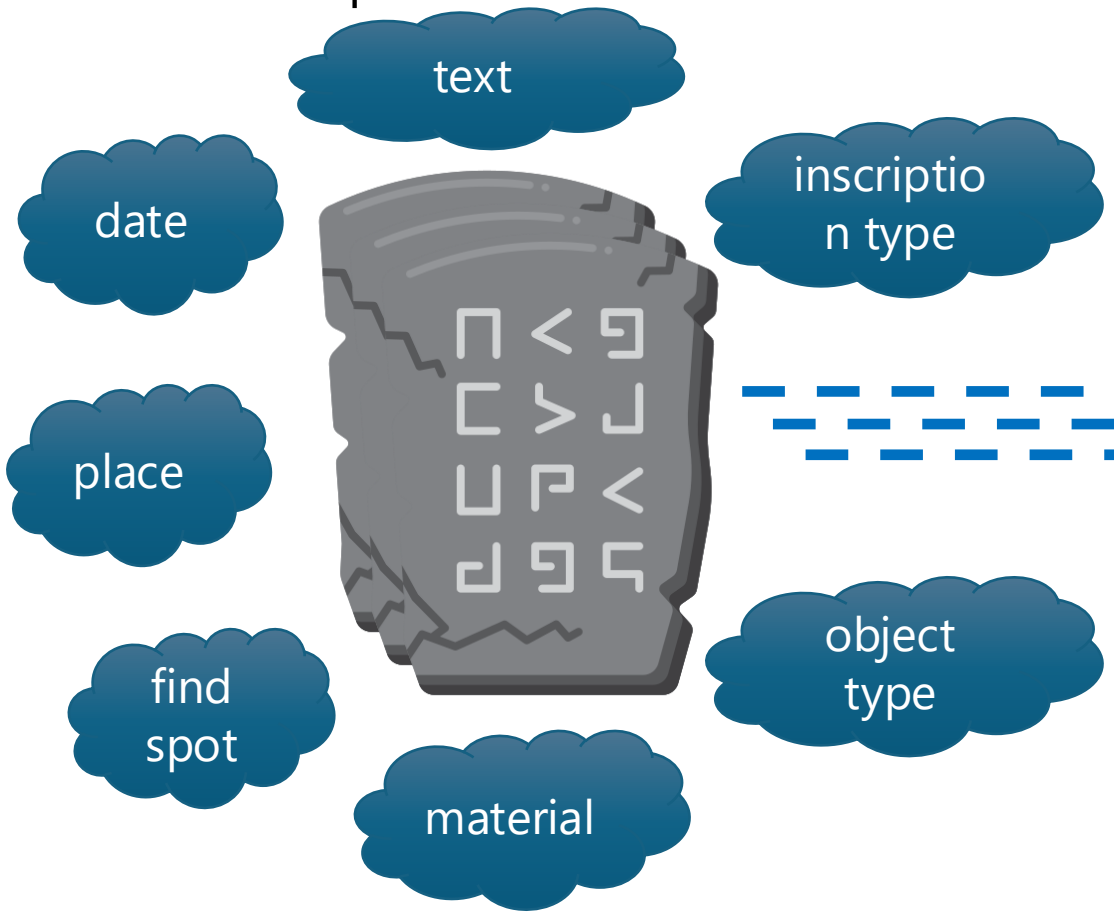
- Research data often originates from heterogeneous sources
- It includes texts, images, measurements, and annotations
- Data is frequently created in discipline-specific formats
- Many datasets are not originally designed for reuse
- The structure of data is often implicit rather than explicit
- This makes integration and comparison difficult
- Therefore, additional processing is required to make data usable



Research data repository

DOCX in Humanities Research

- Focus on presentation



Sylvia Melzer, S. Schiff, F. Weise, K. Harter, and R. Möller, Databasing on demand for research data repositories explained with a large epidoc dataset, Book of Industry Papers, Poster Papers and Abstracts of the CENTERIS 2022 - Conference on ENTERprise Information Systems / ProjMAN 2022 - International Conference on Project MANagement / HCist 2022 - International Conference on Health and Social Care Info , pp. 150--153, 2022. SciKA, Portugal.

DOCX files

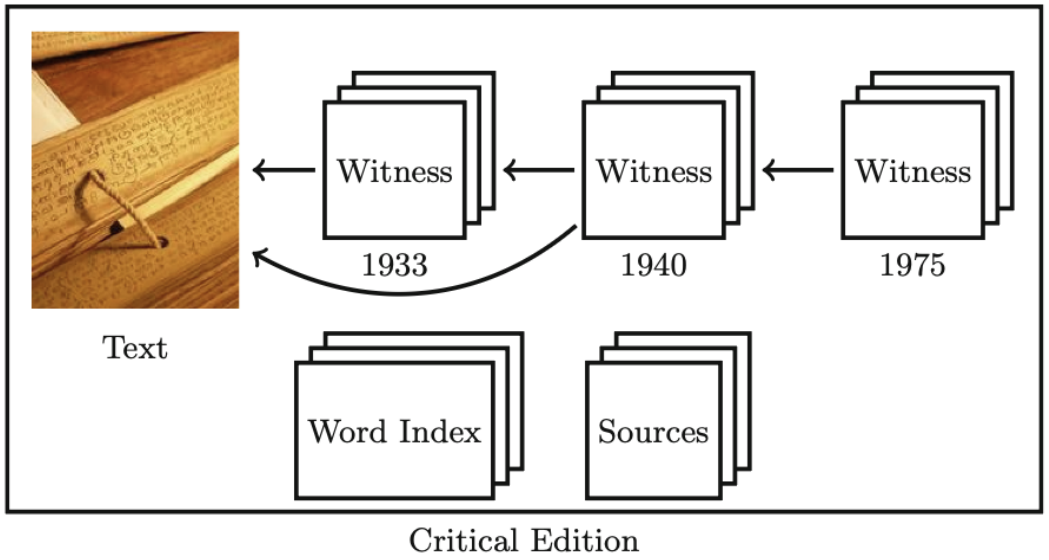
exchange of data and search for information is a challenge

Problem with DOCX

- Focus on visual presentation (WYSIWYG) rather than structure
- Lack of semantic encoding (e.g., persons, places, variants not explicitly defined)
- Data stored in different formats and tools across researchers
- Merging and harmonizing data is time-consuming and error-prone
- Creation of information based on research date (e.g., critical editions) takes years and involves high manual effort
- Difficult to produce digital, FAIR-compliant editions in addition to print
- Limited support for:
  - Automated processing
  - Large-scale analysis
  - Advanced search (e.g., faceted search)
- Result: Data exchange and information retrieval become a major challenge

Critical Editions as Example

- Critical editions are created over many years of research
- They combine multiple textual witnesses and sources
- Scholars work with different tools and formats
- Data integration becomes complex and time-consuming
- The final output is often a printed edition
- Digital reuse is not inherently supported
- This highlights the need for structured representations



Simon Schiff, Sylvia Melzer, Eva Wilden, and Ralf Möller, **TEI-based Interactive Critical Editions**, in *15th IAPR International Workshop on Document Analysis Systems* , Springer, 2022.

Need for Digital and FAIR Data

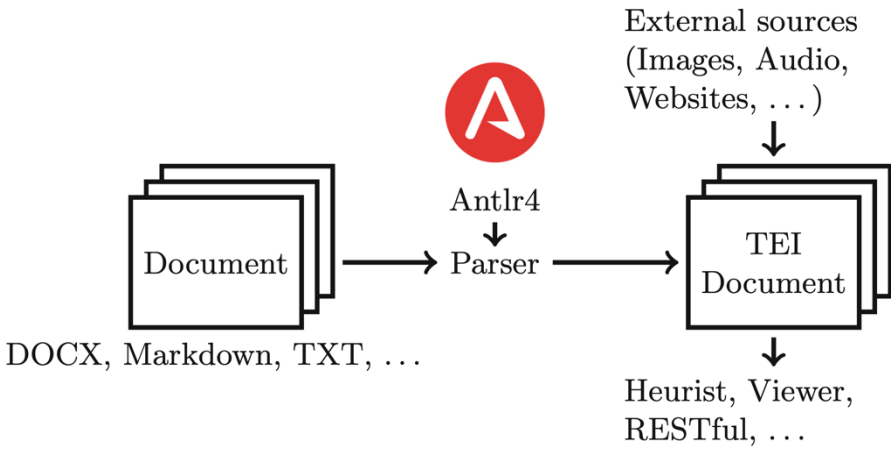
- There is growing demand for digital editions
- Digital data must be searchable and accessible
- Linking between data sources becomes important
- Data should follow FAIR principles
- Reusability is essential for future research
- Machine-readable formats enable automation
- Structured data supports advanced analysis methods

The EASE-Oriented Approach

- **E**xplorable, **A**ccessible in terms of data quality, **S**eeable in operation in new contexts, and **E**asily checkable for reuse (EASE)
- The EASE-oriented approach focuses on supporting existing research practices
- It does not require scholars to abandon their preferred tools
- Instead, it introduces lightweight integration and transformation mechanisms, the goal is to reduce technical barriers e.g., for humanities researchers
- Data is transformed in the background without disrupting workflows
- This enables gradual adoption of structured data practices
- The approach emphasizes usability, flexibility, and sustainability

EASE in Practice (Project: Netamil)

- Documents are created in familiar environments such as Word
- A transformation layer converts them into TEI automatically
- The transformation can be supported by parsers and rules
- Structured data is stored in repositories
- Data can be enriched with external resources
- Outputs can be generated for print and digital formats
- The process remains transparent to the researcher

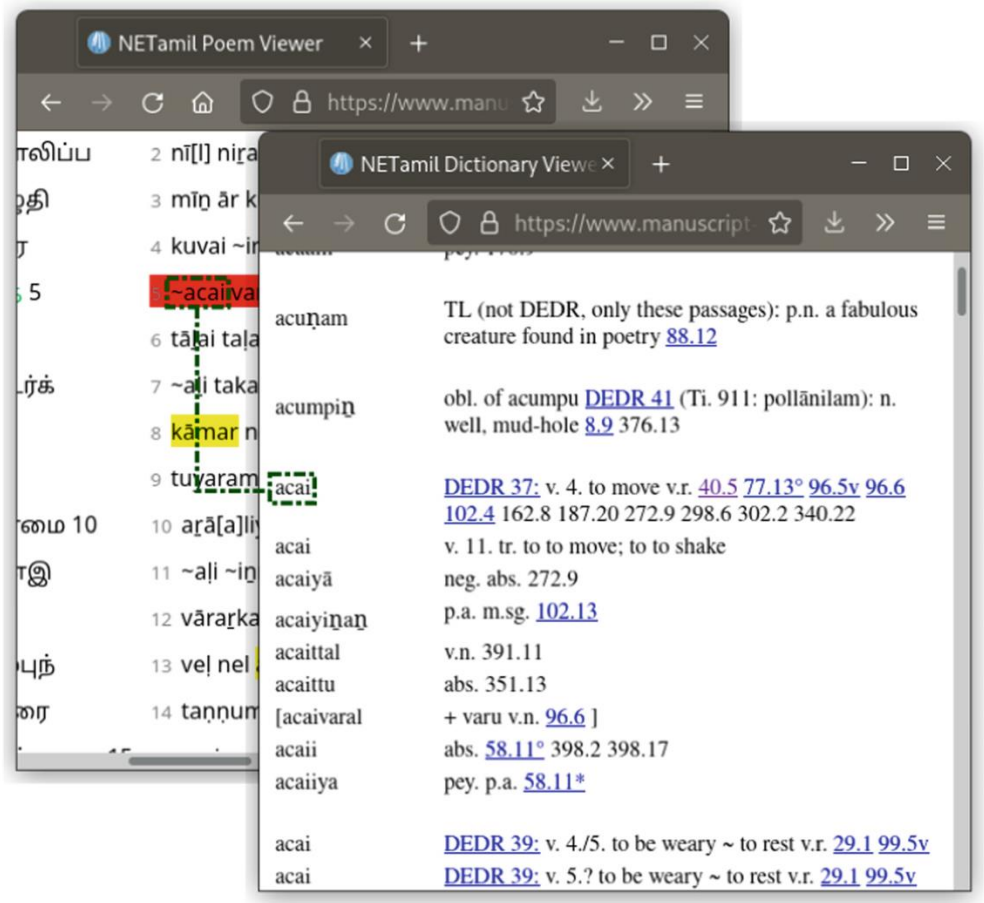


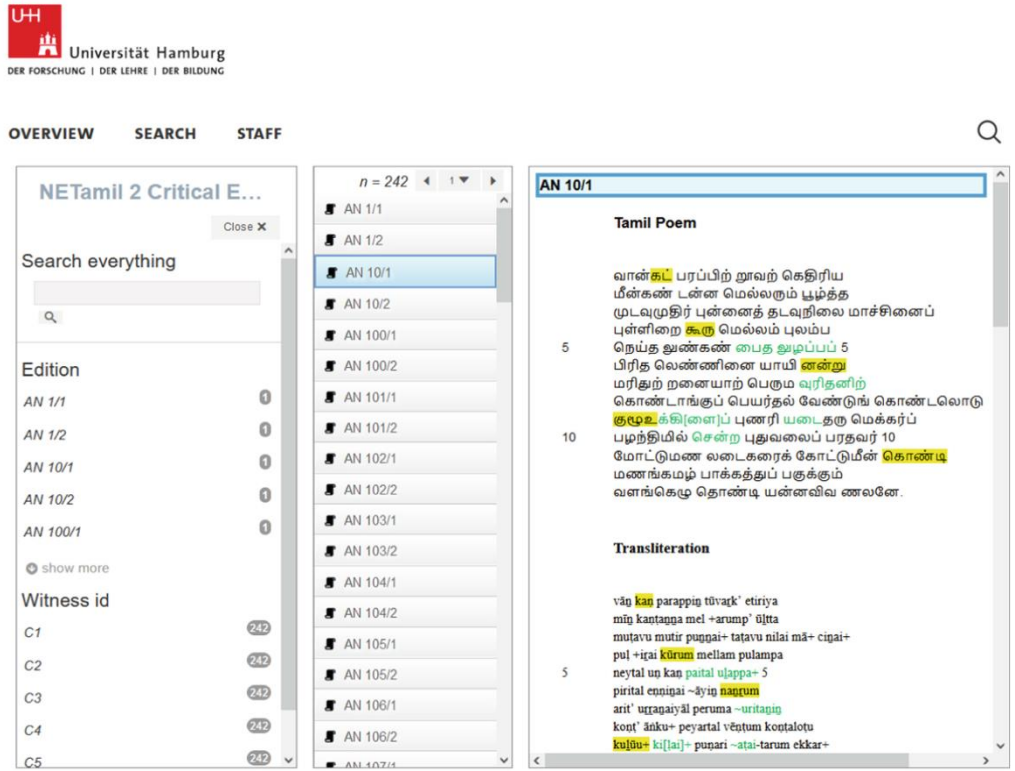
Simon Schiff, Sylvia Melzer, Eva Wilden, and Ralf Möller, **TEI-based Interactive Critical Editions**, in *15th IAPR International Workshop on Document Analysis Systems* , Springer, 2022.

EASE and Data Modeling

- EASE enables the transition from implicit to explicit data models (e.g., DOCX → TEI)
- TEI acts as the intermediate structured model
- Data models emerge from transformation processes
- Researchers can analyze and refine these models
- The approach supports iterative improvement of data quality
- It connects representation with modeling practice.

EASE and Data Modeling





Simon Schiff, Sylvia Melzer, Eva Wilden, and Ralf Möller, **TEI-based Interactive Critical Editions**, in *15th IAPR International Workshop on Document Analysis Systems* , Springer, 2022.

Key Insight of EASE

The goal is not to replace tools, but to extend them.

Structure is added without changing existing workflows.

Data becomes machine-readable without extra burden.

Transformation enables interoperability and reuse.

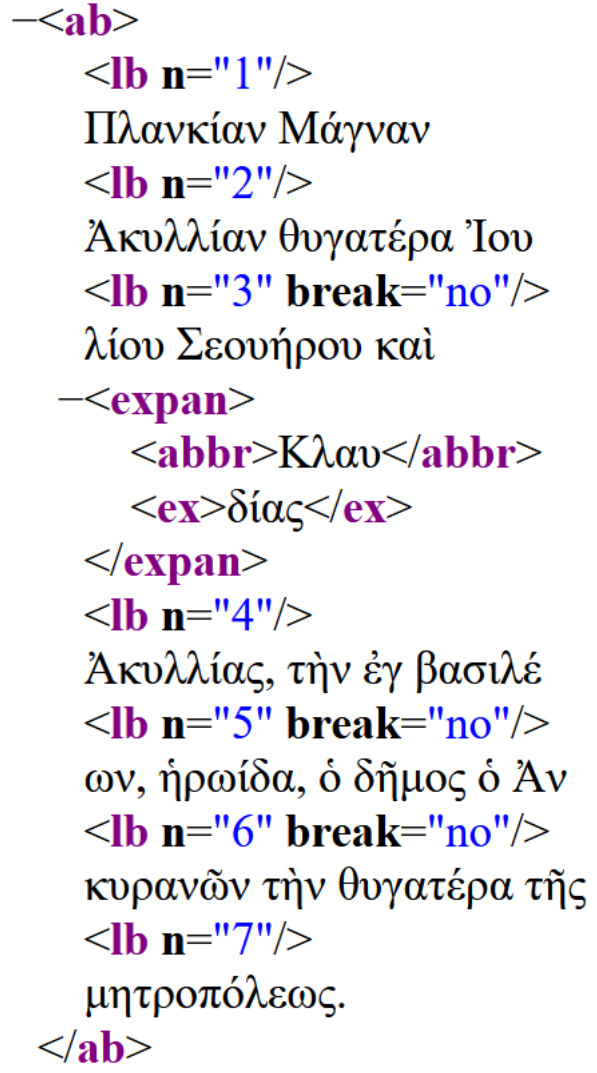
Modeling becomes part of everyday research practice.

This lowers the barrier to digital methods.

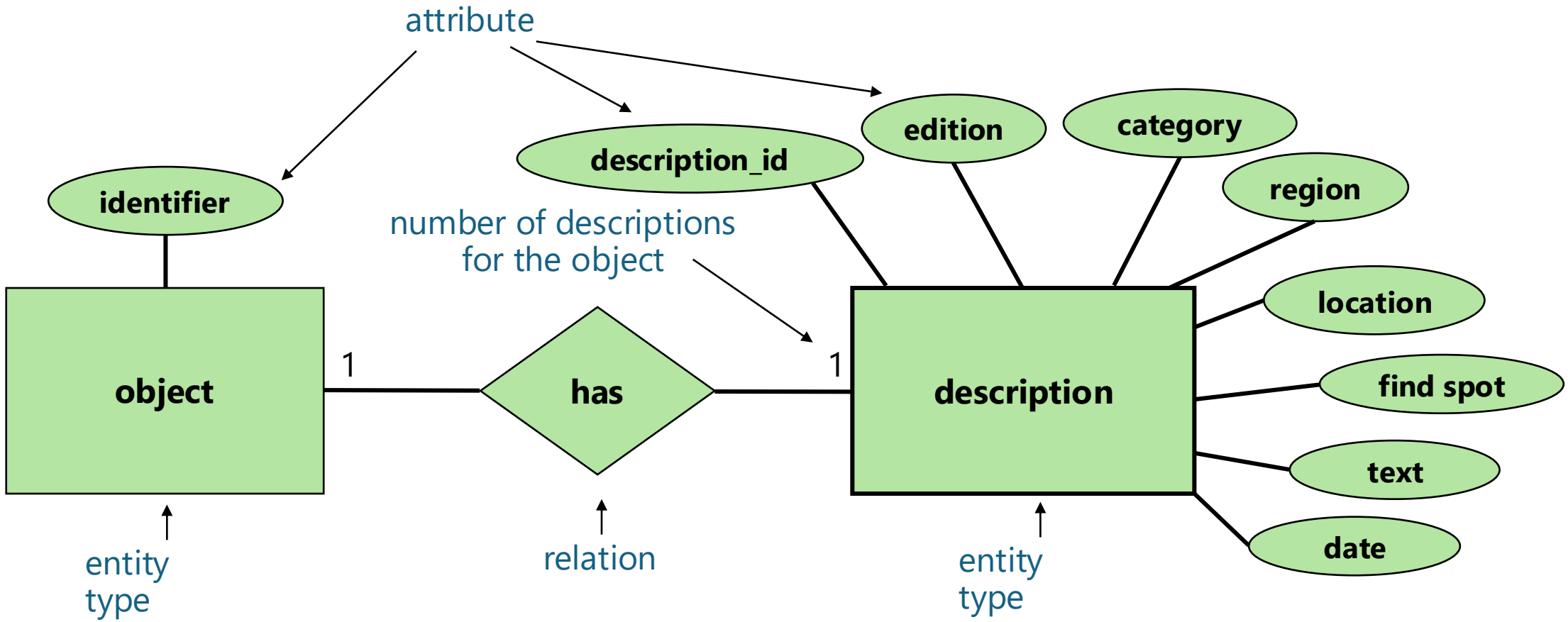
EASE enables sustainable research data ecosystems.

Representation = Modeling

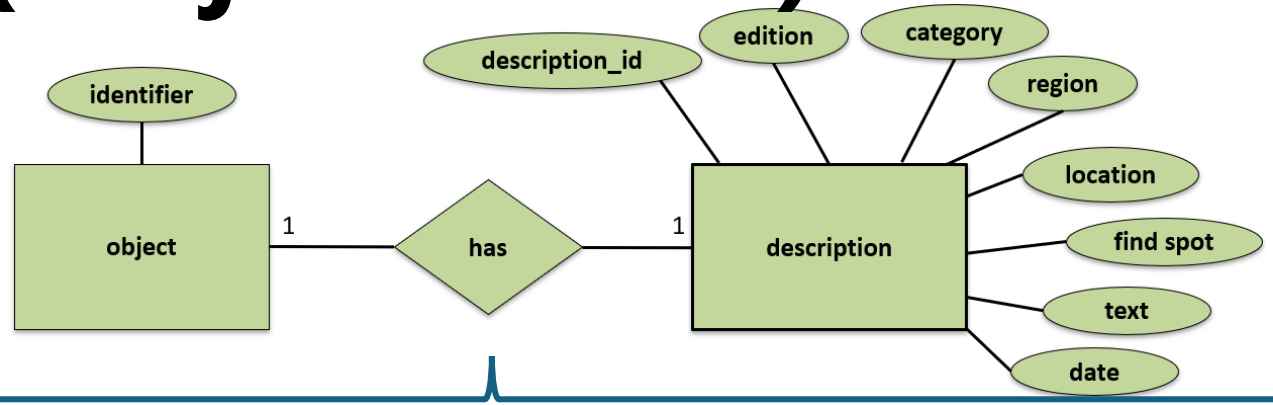
- Every data representation encodes a data model
- A data model is an abstract representation of data
- A format defines what can be expressed in data
- It determines how information is structured
- It also defines relationships between elements
- Representation is therefore not neutral
- It reflects assumptions about the data
- Understanding this is central to data modeling



Data Model (Project: EDAK)



Data Model (Project: EDAK)



- The table represents the **has**-relation between the **object** and the **description**
- The attributes are the column names
- A data set is in one row

Identifier	Description_id	Edition	Category	Region	Location	Find spot	Text	Date
object_id	00002670	MAMA I, Nr. 319	Epigramm	Lykaonien	Laodikeia	Gözlü	Ἀντίοκος {Ἀντίοχος} τόδε σῆμα φί- λῳ περικαλλέει πεδὶ τε[ῦ]- ξεν Ἀρμένιου πόθω [χ]άριν θέτο τ[ῷδ' ἐπὶ τύμβῳ]. ἀντὶ φιλοστοργ[ί]ης [τ]ε τίτλο[ν] ἀμφὶ δ' ἄρ' αὐτῷ πολλὰ κινυρ[ό]- μενος κατηφιάας κὲ προ[—]	

Data from https://www.epigraphik.uni-hamburg.de

Data Model

- Data models ensure consistency across datasets
- They reduce redundancy and duplication
- They support efficient data retrieval
- They improve data quality and reliability
- They enable interoperability between systems
- They make data easier to understand
- They are essential for scalable analysis

Structured vs Unstructured Data

- Unstructured data consists of free-form text or media
- Structured data follows predefined schemas
- Semi-structured data lies between both extremes
- Structured data is easier to query and analyze
- Unstructured data requires preprocessing
- Representation formats influence structure
- Modeling transforms unstructured into structured data

```
<text>
    <body>
        <div n="AN_1/1" type="tamil" xml:lang="ta">
            <head>Tamil Poem</head>
            <l n="1">வண்டுபடத் ததைந்தகண்ணியொண்கழ</l>
            <l n="2">லுருவக் குதிரைமழவரோட்டிய</l>
            <l n="3">முருக னற்போர்நெடுவேளாவி</l>
            <l n="4">யறுகோட் டியானைப்<span style="color:#00B050;">
                பொதுளிய</span>யாங்கட்</l>
            <l n="5">சிறுகா ரோடன்பயிலொடுசேர்த்திய5</l>
            <l n="6">கற்போற் பிரியலமென்றசொற்றா</l>
            <l n="7">மறந்தனர் கொல்லோதோழிசிறந்த</l>
            <l n="8">வேய்மருள் பணைத்தோ<span style="color:#00B050;">
```

Palm leaves with text in Singhalese from Sri Lanka (ca. mid-20th century). Private collection.

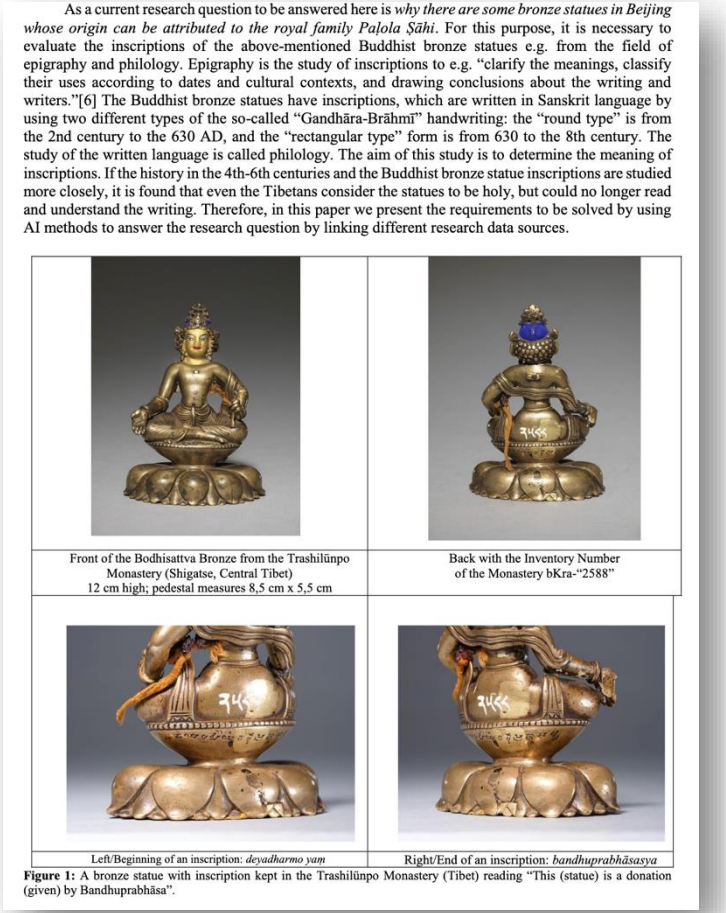
Case Study – DOCX (Weak Model)

- DOCX focuses on visual layout and formatting
- Structure is implicit rather than explicit
- Formatting conveys meaning indirectly
- Machines cannot reliably interpret formattin
- Data extraction becomes complex
- There is no enforced structure
- This results in weak data modeling

In humanities projects, work in Word is also highly structured, though this isn't always immediately apparent. However, in these cases, only the author is aware of the structure. We will take a closer look at how to make this structure explicit.

Implicit Data Model

- Entities are not formally defined
- Relationships are not explicitly encoded
- Data meaning depends on human interpretation
- Structure varies across documents
- No standardization is enforced
- Data integration is difficult
- This limits reuse and analysis



As a current research question to be answered here is *why there are some bronze statues in Beijing whose origin can be attributed to the royal family Paḷola Śāhi*. For this purpose, it is necessary to evaluate the inscriptions of the above-mentioned Buddhist bronze statues e.g. from the field of epigraphy and philology. Epigraphy is the study of inscriptions to e.g. "clarify the meanings, classify their uses according to dates and cultural contexts, and drawing conclusions about the writing and writers."[6] The Buddhist bronze statues have inscriptions, which are written in Sanskrit language by using two different types of the so-called "Gandhāra-Brāhmī" handwriting: the "round type" is from the 2nd century to the 630 AD, and the "rectangular type" form is from 630 to the 8th century. The study of the written language is called philology. The aim of this study is to determine the meaning of inscriptions. If the history in the 4th-6th centuries and the Buddhist bronze statue inscriptions are studied more closely, it is found that even the Tibetans consider the statues to be holy, but could no longer read and understand the writing. Therefore, in this paper we present the requirements to be solved by using AI methods to answer the research question by linking different research data sources.

Front of the Bodhisattva Bronze from the Trashilünpo Monastery (Shigatse, Central Tibet)
12 cm high; pedestal measures 8,5 cm x 5,5 cm

Back with the Inventory Number of the Monastery bKra-"2588"

Left/Beginning of an inscription: *deyadharmo yaṃ*

Right/End of an inscription: *bandhuprabhāsasya*

Figure 1: A bronze statue with inscription kept in the Trashilünpo Monastery (Tibet) reading "This (statue) is a donation (given) by Bandhuprabhāsa".

Hu-von Hinüber, H., & Melzer, S. (2022). On the Awakening of the Buddhological Epigraphy and Philology from the AI. In *Proceedings of the 2nd Workshop on Humanities-Centred Artificial Intelligence (CHAI 2022)* (Band 3301). CEUR-WS.org. https://ceur-ws.org/Vol-3301/paper5.pdf

Consequences

- Data processing requires manual effort
- Automated workflows are limited
- Errors occur during data integration
- Redundancies are common
- Queries cannot be performed efficiently
- Large-scale analysis is impractical

→ **This motivates structured approaches**

Case Study – TEI (Explicit Model)

- The Text Encoding Initiative (TEI) is a collaborative consortium focused on developing and maintaining a standard for digitally representing texts
- Its primary output is a set of Guidelines outlining encoding methods primarily used in the humanities, social sciences, and linguistics
- Since 1994, these Guidelines have been widely adopted by libraries, museums, publishers, and scholars for presenting texts online, supporting research, education, and preservation efforts
- In addition to the Guidelines, the TEI Consortium offers various resources, training events, project showcases, and a bibliography to support learning and usage of TEI
- Giudelines: https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html

TEI: Overview

Front Matter

Title
- i. Releases of the TEI Guidelines
- ii. Dedication
- iii. Preface and Acknowledgments
- iv. About These Guidelines
- v. A Gentle Introduction to XML
- vi. Languages and Character Sets

Back Matter

- Appendix A Model Classes
- Appendix B Attribute Classes
- Appendix C Elements
- Appendix D Attributes
- Appendix E Datatypes and Other Macros
- Appendix F Bibliography
- Appendix G Deprecations
- Appendix H Prefatory Notes

Appendix I Colophon

Text Body

1. The TEI Infrastructure
2. The TEI Header
3. Elements Available in All TEI Documents
4. Default Text Structure
5. Characters, Glyphs, and Writing Modes
6. Verse
7. Performance Texts
8. Transcriptions of Speech
9. Computer-mediated Communication
10. Dictionaries
11. Manuscript Description
12. Representation of Primary Sources
13. Critical Apparatus
14. Names, Dates, People, and Places
15. Tables, Formulæ, Graphics, and Notated Music
16. Language Corpora
17. Linking, Segmentation, and Alignment
18. Simple Analytic Mechanisms
19. Feature Structures
20. Graphs, Networks, and Trees
21. Non-hierarchical Structures
22. Certainty, Precision, and Responsibility
23. Documentation Elements
24. Using the TEI

TEI: Header

```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>
<!-- title of the resource -->
   </title>
  </titleStmt>
  <publicationStmt>
   <p>
<!-- Information about distribution of the resource -->
   </p>
  </publicationStmt>
  <sourceDesc>
   <p>
<!-- Information about source from which the resource derives -->
   </p>
  </sourceDesc>
 </fileDesc>
</teiHeader>
```

https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html

TEI: Dictionary

```
<body>
 <div>
  <head>English-French</head>
  <entry>
   <form>
    <orth>cat</orth>
   </form>
<!-- ... -->
  </entry>
  <entry>
   <form>
    <orth>dog</orth>
   </form>
<!-- ... -->
  </entry>
  <entry>
   <form>
    <orth>horse</orth>
   </form>
<!-- ... -->
  </entry>
 </div>
 <div>
  <head>French-English</head>
  <entry>
   <form>
    <orth>chat</orth>
   </form>
<!-- ... -->
  </entry>
  <entry>
   <form>
    <orth>chien</orth>
   </form>
<!-- ... -->
  </entry>
  <entry>
   <form>
    <orth>cheval</orth>
   </form>
<!-- ... -->
  </entry>
 </div>
</body>
```

Example: Netamil project

```
1 <l n="1">
2     <entry>
3         <form>
4             <orth>acai</orth>
5         </form>
6         <def>DEDR 37: v. 4. to move v.r. 40.5 77.13° 96.5v 96.6 102.4 162.8 187.20
              272.9 298.6 302.2 340.22</def>
7     </entry>
8 </l>
```

- DEDR 37 refers to entry 37 in the “A Dravidian etymological dictionary” (DEDR: https://dsal.uchicago.edu/dictionaries/burrow/)
- v. 4. indicates the 4th verb in DEDR
- to move is the English translation
- v.r. means verbal root
- series of numbers: 40.5, 77.13°, 96.5v, 96.6, 102.4, 162.8, 187.20, 272.9,
- 298.6, 302.2, 340.22 shows in which poems (number before .) the word occurs and in which line (number after .)
- “°” means primary variant second strand and “v” means secondary variant

TEI

TEI as Data Model

- TEI encodes entities explicitly
- It defines relationships within texts
- It structures documents hierarchically
- It integrates “metadata” and content
- It enables machine-readable representation
- It supports consistent data encoding
- It acts as a formal data model

TEI Structure

- TEI documents have a standardized structure
- The teiHeader contains metadata information
- The text element contains the content
- Elements define structural units
- Attributes provide additional detail
- Hierarchies reflect document organization
- This structure enables processing

TEI

Advantages

- TEI produces machine-readable data
- It supports interoperability between systems
- It enables long-term data preservation
- It facilitates data reuse
- It supports automated processing
- It enables advanced querying
- It improves research efficiency

Disadvantages

- TEI documents can be complex and difficult to understand for non-experts
- The flexibility of TEI may lead to inconsistent encoding across projects
- Interoperability can be limited when projects use different TEI customizations
- TEI files are not always easily discoverable without proper metadata infrastructure
- Machine readability is high, but practical reuse often requires additional processing
- Long-term accessibility depends on tooling, documentation, and community support

EpiDoc (Epigraphic Documents in TEI XML)

- EpiDoc is a collaborative international initiative that offers guidelines and tools for encoding scholarly and educational editions of ancient texts
- It employs a specific subset of TEI's standards for digitally representing texts and was originally designed for the publication of digital editions of ancient inscriptions
- Its application has broadened to encompass the publication of papyri and manuscripts, as demonstrated by platforms like Papyri.info
- EpiDoc covers not only the transcription and editorial aspects of the texts but also the historical context and materials of the artifacts on which these texts are found, including manuscripts, monuments, tablets, papyri, and other objects containing text
- Guidelines: https://epidoc.stoa.org/gl/latest/

Case Study – EpiDoc (Explicit Model) – Project EDAK

```
<TEI xml:lang="en">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>SEG 6, 29</title>
      </titleStmt>
      <publicationStmt>
        <authority>Epigraphische Datenbank zum antiken Kleinasien – Universität Hamburg</authority>
        <idno type="localID">EDAK00001210</idno>
        <idno type="filename">EDAK00001210</idno>
        <availability>[...]</availability>
      </publicationStmt>
      <sourceDesc>
        <msDesc>
          <msIdentifier/>
          <msContents>[...]</msContents>
          <physDesc>[...]</physDesc>
          <history>
            <origin>
              <origPlace>
                <settlement ref="origPlace.xml#fundort__83"/>
              </origPlace>
              <origDate notBefore="0301" notAfter="0400">4 Jh.
```

```
<body>
  <div type="edition" xml:lang="gr">
    <ab>
      <lb n="1"/><supplied reason="lost">  </supplied>
      <lb n="2" break="no"/>  <expan><abbr>Θ<ex>  </ex> </abbr></expan>
      <lb n="3"/>Θ
      <lb n="4" break="no"/>  ,<expan><abbr>  </abbr><ex>  </ex></expan> ',
      <lb n="5"/><expan><abbr>  </abbr><ex>  </ex></expan>M  '.
    </ab>
  </div>
  <div type="commentary">
    <p>Grabinschrift für Theodoros; gef. in Ankara, einmal im Augustus–Tempel ( bis 1927).</p>
  </div>
  <div type="bibliography">
    <listBibl>
      <bibl>
        <ptr target="EDAK_bibliography.xml#SEG"/>
        <citedRange>6, 29 (zur Datierung)</citedRange>
      </bibl>
      <bibl>
        <ptr target="EDAK_bibliography.xml
```

```
7         <publicationStmt>
8            <authority>Epigraphische Datenbank
                zum antiken Kleinasien –
                Universität Hamburg</authority
                >
9            <idno type="localID">EDAK00001210
                </idno>
10           <idno type="filename">EDAK00001210
                </idno>
11           <availability>[...]</availability>
12        </publicationStmt>
13        <sourceDesc>
14           <msDesc>
15              <msIdentifier/>
16              <msContents>[...]</msContents>
17              <physDesc>[...]</physDesc>
18              <history>
19                 <origin>
20                    <origPlace>
21                       <settlement ref="
                            origPlace.xml#
                            fundort_83"/>
22                    </origPlace>
23                    <origDate notBefore="0301"
                         notAfter="0400">4 Jh.
                         n.Chr. (Hondius)</
                         origDate>
24                 </origin>
25                 <provenance type="found">
26                    <placeName ref="findspot.
                         xml#fundort_83"/>
27                 </provenance>
28              </history>
29           </msDesc>
30        </sourceDesc>
31     </fileDesc>
32     <encodingDesc>[...]</encodingDesc>
33     <profileDesc>[...]</profileDesc>
34     <revisionDesc>[...]</revisionDesc>
35  </teiHeader>
```

```
                  <abbr>Θ<ex> </ex> </abbr></
                  expan>
6           <lb n="3"/>Θ
7           <lb n="4" break="no"/>     ,<expan><
                  abbr>  </abbr><ex>     </ex></
                  expan> ',
8           <lb n="5"/><expan><abbr> </abbr><
                  ex> </ex></expan>M    '.
9        </ab>
10    </div>
11    <div type="commentary">
12       <p>Grabinschrift für Theodoros; gef. in
               Ankara, einmal im Augustus–Tempel (
               bis 1927).</p>
13    </div>
14    <div type="bibliography">
15       <listBibl>
16          <bibl>
17             <ptr target="EDAK_bibliography.xml
                     #SEG"/>
18             <citedRange>6, 29 (zur Datierung)</
                     citedRange>
19          </bibl>
20          <bibl>
21             <ptr target="EDAK_bibliography.xml
                     #Anderson1899"/>
22             <citedRange>97 Nr. 80</citedRange>
23          </bibl>
24          <bibl>
25             <ptr target="EDAK_bibliography.xml
                     #Jerphanion1928"/>
26             <citedRange>–287288 Nr. 61 (zu Z. 1,
                     –45).</citedRange>
27          </bibl>
28       </listBibl>
29    </div>
30    <div type="translation"> <p/>
31    </div>
32    <div type="apparatus"> <p/>
33    </div>
34 </body>
```

EpiDoc Stylesheets

machine-readable format

```
…
<ab>

<lb n="1"/>
Πλανκίαν Μάγναν

<lb n="2"/>
Ἀκυλλίαν θυγατέρα Ἰου

<lb n="3" break="no" />
λίου Σευήρου καί
<abbr>Κλαυ</abbr>
<ex>δίας</ex>

<lb n="4"/>
Ἀκυλλίας, τὴν ἐν βασιλέ

<lb n="5" break="no" />
ων, ἡρωίδα, ὁ δῆμος ὁ Ἀν

<lb n="6" break="no" />
κυρανῶν τὴν θυγατέρα τῆς

<lb n="7"/>
μητροπόλεως.
</ab>
…
```

EpiDoc stylesheets

human-readable format

1 Πλανκίαν Μάγναν
2 Ἀκυλλίαν θυγατέρα Ἰου-
3 λίου Σευήρου καί Κλαυ(δίας)
4 Ἀκυλλίας, τὴν ἐν βασιλέ-
5 ων, ἡρωίδα, ὁ δῆμος ὁ Ἀν-
6 κυρανῶν τὴν θυγατέρα τῆς
7 μητροπόλεως.

JSON as Data Representation (PATHs Project Example)

- JSON is a lightweight format for representing structured data
- It is widely used in web applications and APIs
- Data is organized as key-value pairs
- It supports nested structures and lists
- It is easy to read for both humans and machines
- JSON is commonly used for data exchange between systems
- It provides a flexible alternative to XML-based representations


```
{
    "head": {
        "shortsql": "@manuscripts~?bindings|=|1~-200:0",
        "total_rows":156,
        "total_pages":6,
        "table":"paths__manuscripts",
        "stripped_table":"manuscripts",
        "table_label":"Manuscripts",
        "page":1,
        "no_records_shown":156,
        "fields": {[...]},
    },
    "records": [
        {
            "0": {
                "id": "18",
                "creator": "1",
                "cmclid": "CMCL.AS",
                "tm": "108136",
                "ldab": "108136",
                "lcbm": null,
                "dbmnt": null,
                "alias": "Codex Bruce",
                "issinglefrag": "0",
                "isbookbinding": null,
```


https://bdus.cloud/db/api/paths?verb=search&shortsql=@manuscripts~?bindings%7C=%7C1~-200:0

Bookbinding Data in JSON

- The example shows research data from the PATHs project
- Each record represents one bookbinding entity
- Attributes include stratigraphy, modern history, and book form
- Data is organized into fields and records
- Each record is identified by an internal identifier
- Textual descriptions are stored as string values
- The structure reflects a simplified data model


```
{
   "head": {
      "shortsql": "@manuscripts~?bindings|=|1~-200:0",
      "total_rows":156,
      "total_pages":6,
      "table":"paths__manuscripts",
      "stripped_table":"manuscripts",
      "table_label":"Manuscripts",
      "page":1,
      "no_records_shown":156,
      "fields": {[...]},
   },
   "records": [
      {
         "0": {
            "id": "18",
            "creator": "1",
            "cmclid": "CMCL.AS",
            "tm": "108136",
            "ldab": "108136",
            "lcbm": null,
            "dbmnt": null,
            "alias": "Codex Bruce",
            "issinglefrag": "0",
            "isbookbinding": null,
```


https://bdus.cloud/db/api/paths?verb=search&shortsql=@manuscripts~?bindings%7C=%7C1~-200:0

Audition Certificate Data in JSON


```
{
  "text": "[...] لضافلا ملا علا هيقفلا هبحا صباتكلا اذ هعيمجيًّلععمسهللا همحرفلؤملا طخبعامسة قبطة روص",
  "properties": [
    {
      "index": 1,
      "guid": "b8fdb5ba-fcd8-4b8d-b652-29a2bb5066f9",
      "type": "place",
      "text": "قشمد"",
      "startIndex": 419,
      "endIndex": 422,
      "attributes": {
        "location": [ "bilad-al-sham", "damascus"]
      },
      "isZeroPoint": false
    },
    {
```

```
5          "index": 1,
6          "guid": "b8fdb5ba-fcd8-4b8d-b652-29a2bb5066f9",
7          "type": "place",
8          "text": قشمد"",
9          "startIndex": 419,
10         "endIndex": 422,
11         "attributes": {
12           "location": [ "bilad-al-sham", "damascus"]
13         },
14         "isZeroPoint": false
15       },
16       {
17         "index": 2,
18         "guid": "1a933be2-c462-4a36-a386-217a4dccf6ce",
19         "type": "person",
20         "text": "باهش
               يعفاشلايقشمدلايراصنألاناوعجنبسابعنبدمحمهللادبعيبأنيدلالامكلجألاخيشلانبدمحأسابعلاوبأنيدلا",
21         "startIndex": 107,
22         "endIndex": 218,
23         "attributes": {
24           "role": ["reader", "owner"],
25   [...]
```

JSON vs TEI

JSON

- Typically flatter and more application-oriented
- Easier to process in programming environments
- Commonly used for APIs and data exchange

TEI

- Richer and more expressive for textual data
- Captures detailed semantics and annotations
- Is used for scholarly encoding and preservation

Both formats represent data models in different ways

JSON as a Data Model

- JSON structures represent entities as objects
- Keys correspond to attributes of these entities
- Nested objects represent relationships between entities
- Arrays represent collections of entities
- Identifiers are required to link related data
- The structure reflects modeling decisions made during transformation
- JSON therefore encodes a data model, not just a format

TEI and JSON as Alternative Representations

- The same data can be represented in TEI or JSON
- TEI uses hierarchical XML structures
- JSON uses object-based structures
- Both encode entities, attributes, and relationships
- TEI focuses on semantic richness and textual detail
- JSON focuses on usability in applications and APIs
- Transformation between both formats is possible

Multiple Representations

- The same research data can be represented in different formats
- DOCX represents data implicitly
- TEI and EpiDoc represent data explicitly and semantically
- JSON represents data in a structured and flexible way
- Each format encodes a different data model
- The choice of format affects analysis possibilities
- Understanding representations is key to data modeling

Schema & Validation (Relax NG)

Why Validation?

- Data must follow consistent rules
- Errors must be detected early
- Structure must be enforced
- Validation ensures correctness
- It supports interoperability
- It improves data quality
- It enables reliable processing

Relax NG

- Elements may be added in an TEI/EpiDoc document in a standardized way with the RELAX NG standard
- RELAX NG is a schema language for XML
- It is also an International Standard ISO/IEC 19757-2 and is part of ISO/IEC 19757 DSDL (Document Schema Definition Languages)
- RELAX NG has an XML syntax
- https://relaxng.org/

```
<?xml version="1.0" encoding="utf-8"?>
<grammar xmlns:xlink="http://www.w3.org/1999/xlink"
         xmlns:tei="http://www.tei-c.org/ns/1.0"
         xmlns:teix="http://www.tei-c.org/ns/Examples"
         xmlns="http://relaxng.org/ns/structure/1.0"
         datatypeLibrary="http://www.w3.org/2001/XMLSchema-datatypes"
         ns="http://www.tei-c.org/ns/1.0">
<define name="tei_model.pPart.msdesc">
    <choice>
        <ref name="tei_catchwords"/>
        <ref name="tei_dimensions"/>
        <ref name="tei_heraldry"/>
        <ref name="tei_locus"/>
        <ref name="tei_locusGrp"/>
        <ref name="tei_material"/>
        <ref name="tei_objectType"/>
        <ref name="tei_origDate"/>
        <ref name="tei_origPlace"/>
        <ref name="tei_secFol"/>
        <ref name="tei_signatures"/>
        <ref name="tei_stamp"/>
        <ref name="tei_watermark"/>
    </choice>
</define>
<define name="tei_title">
    <element name="title">
        <a:documentation xmlns:a="http://relaxng.org/ns/compatibility/annotations
            /1.0">(title) contains a title for any kind of work. [3.12.2.2. Titles,
            Authors, and Editors 2.2.1. The Title Statement 2.2.5. The Series
            Statement]</a:documentation>
        <ref name="tei_macro.paraContent"/>
        <ref name="tei_att.global.attributes"/>
```

Relax NG

- Elements may be added in an TEI/EpiDoc document in a standardized way with the RELAX NG standard
- RELAX NG is a schema language for XML
- It is also an International Standard ISO/IEC 19757-2 and is part of ISO/IEC 19757 DSDL (Document Schema Definition Languages)
- RELAX NG has an XML syntax
- https://relaxng.org/

```
16          <ref name="tei_objectType"/>
17          <ref name="tei_origDate"/>
18          <ref name="tei_origPlace"/>
19          <ref name="tei_secFol"/>
20          <ref name="tei_signatures"/>
21           <ref name="tei_stamp"/>
22           <ref name="tei_watermark"/>
23       </choice>
24 </define>
25 <define name="tei_title">
26          <element name="title">
27             <a:documentation xmlns:a="http://relaxng.org/ns/compatibility/annotations
                  /1.0">(title) contains a title for any kind of work. [3.12.2.2. Titles,
                  Authors, and Editors 2.2.1. The Title Statement 2.2.5. The Series
                  Statement]</a:documentation>
28             <ref name="tei_macro.paraContent"/>
29             <ref name="tei_att.global.attributes"/>
30             <ref name="tei_att.typed.attribute.subtype"/>
31             <ref name="tei_att.canonical.attributes"/>
32             <ref name="tei_att.datable.attributes"/>
33             <optional>
34                <attribute name="type">
35                   <a:documentation xmlns:a="http://relaxng.org/ns/compatibility/
                        annotations/1.0">classifies the title according to some convenient
                        typology.
36 Sample values include: 1] main; 2] sub (subordinate); 3] alt (alternate); 4] short; 5] desc
      (descriptive)</a:documentation>
37                   <data type="token">
38                      <param name="pattern">[^\p{C}\p{Z}]+</param>
39                   </data>
40                </attribute>
41             </optional>
42             [...]
43          </element>
44       </define>
```

Function of Relax NG

- It defines structural constraints
- It validates XML documents
- It detects invalid elements
- It enforces modeling rules
- It ensures consistency across datasets
- It supports automated workflows
- It strengthens data reliability

Same Standard, Different Data Models (EpiDoc Case)

- Different projects can use the same standard such as EpiDoc
- Examples include CGRN and EDAK, both based on TEI/EpiDoc
- Despite using the same framework, their schemas can differ
- Each project defines its own elements, attributes, and constraints
- Modeling decisions depend on research questions and domain needs
- This leads to variations in structure and interpretation of data
- Therefore, interoperability is not guaranteed by the standard alone

Same Standard, Different Data Models (EpiDoc Case)

```
▼<origin>
  ▼<p>
     <origDate notBefore="-0550" notAfter="-0500">ca. 550-500 BC</origDate>
   </p>
  ▶<p>
   ...
   </p>
 </origin>
```

```
▼<origin>
  ▶<origPlace>
   ...
   </origPlace>
   <origDate notBefore="0101" notAfter="0200">A) 2nd cent. CE (Taeuber, letters)</origDate>
   <origDate notBefore="0301" notAfter="0400">B) 4th cent. CE (Taeuber, letters)</origDate>
 </origin>
```

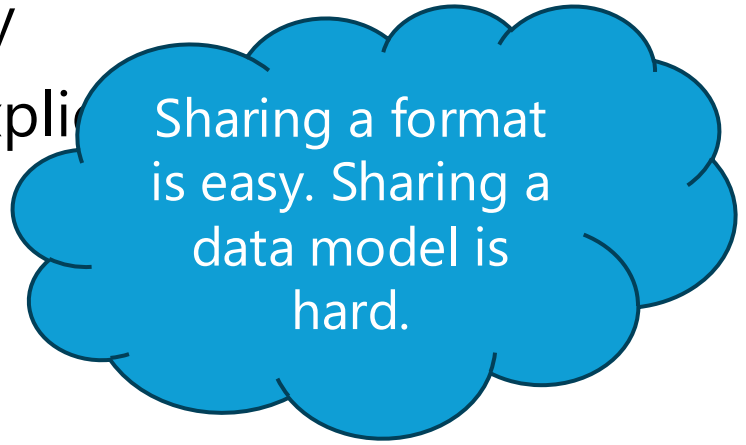
Same Standard, Different Data Models (EpiDoc Case)

```
<provenance>
  <p>
    <placeName type="ancientFindspot" key="Ephesos" n="Asia_Minor_and_Anatolia">
      <ref target="https://pleiades.stoa.org/places/599612" type="external">Ephesos</ref>
    </placeName>


<origPlace>
  <placeName type="ancientFindspot">
    <settlement>
      <ref target="https://pleiades.stoa.org/places/599612/">Ephesos</ref>
      <location>
        <geo>37.9407625 27.340307</geo>
      </location>
    </settlement>
    <region>
      <ref target="https://pleiades.stoa.org/places/550597/ionia/">Ionia</ref>
      <location>
        <geo/>
      </location>
    </region>
  </placeName>
</origPlace>
```

Implications for Data Modeling

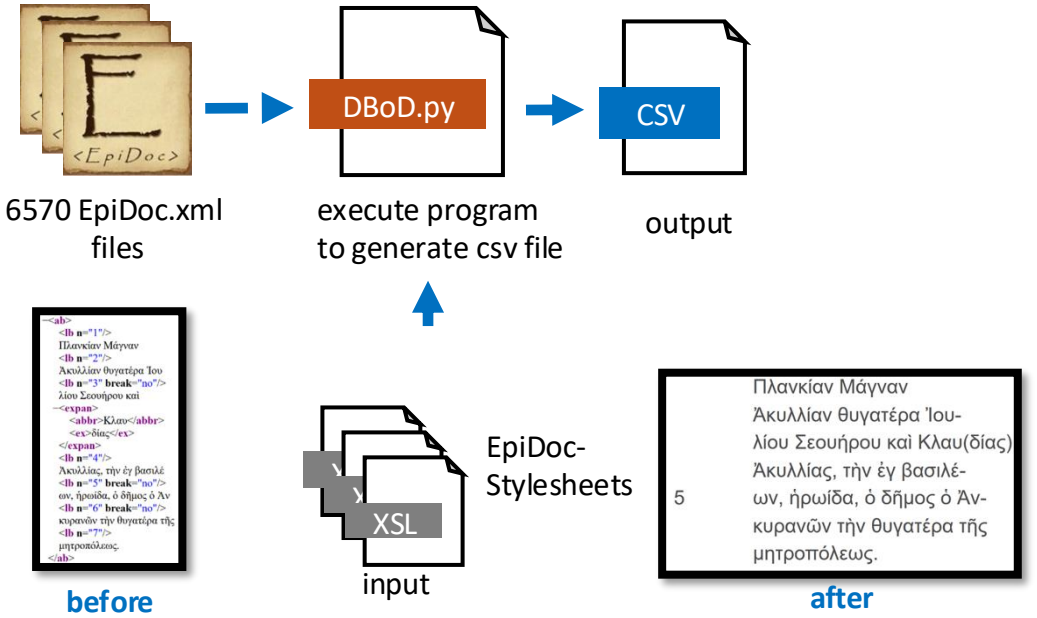
- A standard provides a common vocabulary but not a fixed model
- Customization allows flexibility but introduces heterogeneity
- Data integration requires understanding project-specific schemas
- Validation (e.g., Relax NG) enforces local consistency only
- Cross-project comparison becomes more complex
- Shared conventions are needed for true interoperability
- Modeling decisions must be documented and made expli



From EpiDoc/TEI to Structured Data

Transformation Concept

- EpiDoc and TEI data can be transformed into other formats
- XML structures can be flattened into tables
- Data extraction focuses on entities
- Relationships must be preserved
- Transformation requires modeling decisions
- Tools support automated conversion
- This enables further processing



From EpiDoc/TEI to Structured Data

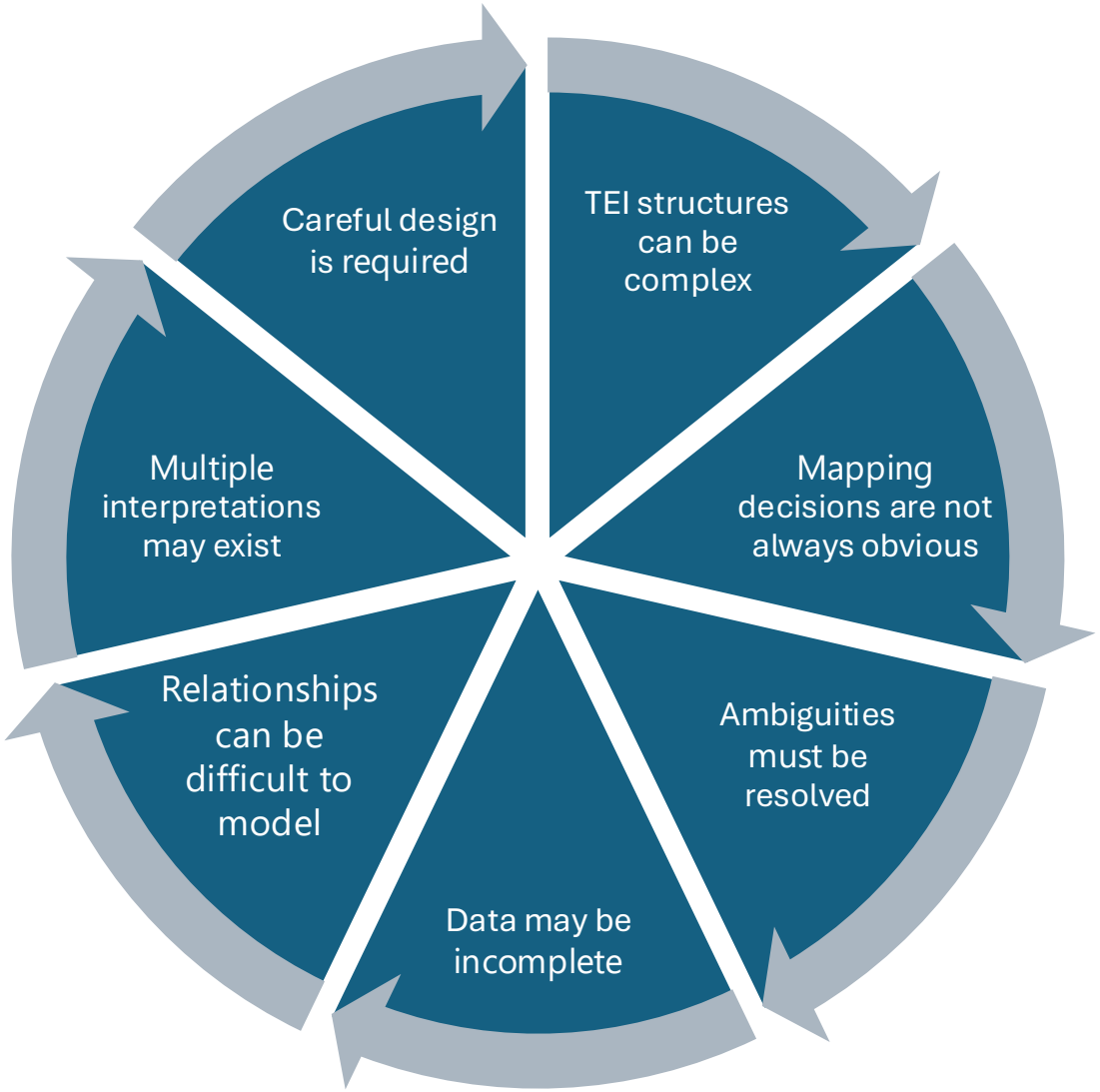
Extracting Data

- Entities such as person names, object type and date can be extracted
- Places can be identified and structured
- Relationships between entities are mapped
- Attributes become table columns
- Identifiers are required for linking
- Data must be cleaned and standardized
- This produces structured datasets

TEI to CSV

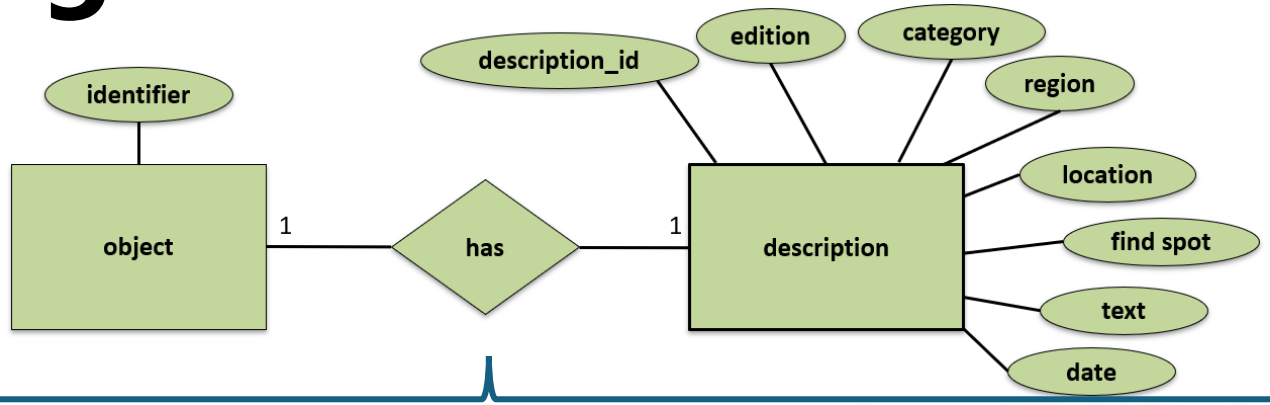
- CSV represents tabular data
- Hierarchical data must be simplified
- Each row represents one entity instance
- Columns represent attributes
- Relationships require additional tables
- Information loss must be managed
- CSV enables database integration

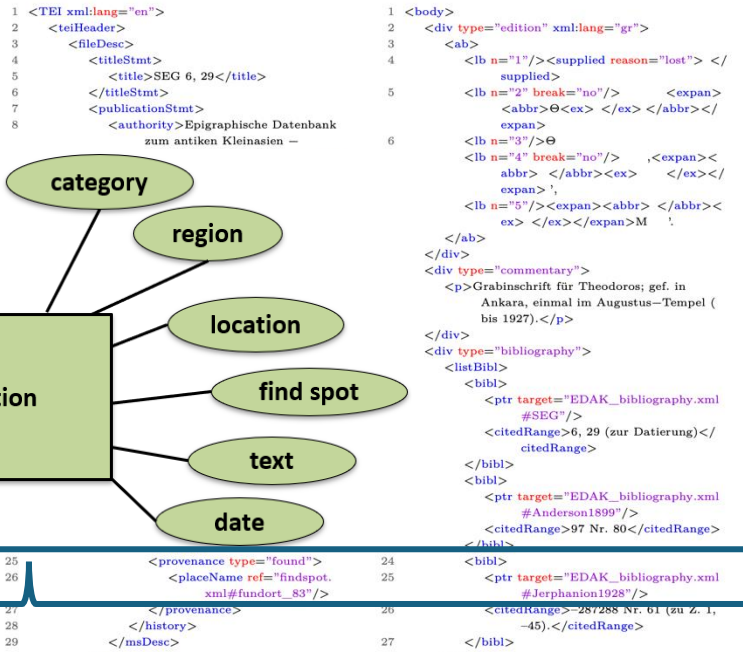
Challenges


Careful design is required
TEI structures can be complex
Mapping decisions are not always obvious
Ambiguities must be resolved
Data may be incomplete
Relationships can be difficult to model
Multiple interpretations may exist

Understanding the Model

- Identify entities within the dataset
- Determine attributes of each entity
- Analyze relationships between entities
- Define keys and identifiers
- Evaluate consistency of data
- Consider normalization aspects
- Reflect on modeling decisions





Identifier	Description_id	Edition	Category	Region	Location	Find spot	Text	Date
object_id	00002670	MAMA I, Nr. 319	Epigramm	Lykaonien	Laodikeia	Gözlü	Ἀντίοκος {Ἀντίοχος} τόδε σῆμα φί- λῳ περικαλλέει πεδὶ τε[ῦ]- ξεν Ἀρμένιου πόθω [χ]άριν θέτο τ[ῷδ' ἐπὶ τύμβῳ]. ἀντὶ φιλοστοργ[ί]ης [τ]ε τίτλο[ν] ἀμφὶ δ' ἄρ' αὐτῷ πολλὰ κινυρ[ό]- μενος κατηφιάας κὲ προ[—]	

Why Normalization Here?

- Data extracted from TEI is often redundant or inconsistent
- CSV representations may duplicate information across rows
- Databases require well-structured data models
- Normalization helps to improve data quality
- It reduces redundancy and avoids inconsistencies
- It prepares data for efficient querying
- It connects transformation with database design

First Normal Form (1NF)

- Each field contains atomic values
- No repeating groups are allowed
- Data is organized in rows and columns
- Each record is uniquely identifiable
- Structure becomes consistent
- Queries become easier

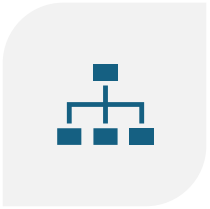
Second Normal Form (2NF)

- Data must already satisfy 1NF
- No partial dependencies are allowed
- Attributes depend on the full key
- Redundancy is reduced
- Data integrity is improved
- Relationships are clarified

Third Normal Form (3NF)

- Data must satisfy 2NF
- No transitive dependencies are allowed
- Attributes depend only on the key
- Redundancy is minimized
- Consistency is improved
- Updates become simpler

TEI vs Relational Models



TEI USES HIERARCHICAL STRUCTURES

DATABASES USE TABULAR STRUCTURES

TRANSFORMATION IS REQUIRED BETWEEN BOTH

TEI CAPTURES RICH SEMANTICS

DATABASES SUPPORT EFFICIENT QUERIES

BOTH COMPLEMENT EACH OTHER

Example Projects

- Projects (EDAK, NETamil, ACP, PATH) differ in domain and scope.
- All require structured data models.
- Interoperability is a common goal.
- These examples illustrate real-world applications.

Conclusion

- Research data must be structured
- Representation formats define models
- TEI and EpiDoc provide rich semantic encoding
- JSON provides structured, application-oriented representation
- Relax NG ensures correctness
- Transformation enables analysis
- Data models improve quality
- Structured data enables FAIR principles

Discussion

- Is TEI a data format, a data model, or both?
- What is lost and what is gained when transforming TEI into CSV or JSON?
- Which representation is best suited for analysis?