DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

# “Learning to Cite” A Framework for the Automatic Construction of Citations


Gianmaria Silvello

Information Management Systems Research Group
Department of Information Engineering
University of Padua

gianmaria.silvello@unipd.it
http://www.dei.unipd.it/~silvello



Seminar/Working Group on Provenance

University of Pennsylvania, 12 February 2016

# Outline

- Motivations and main goal

- XML and digital archives: A use case

- Learning to cite framework

- Experimental evaluation

- Open Questions

# Why Data Citation is Important?

- Give credit to data creators and curators (and institutions)

- Repeatability, reproducibility and generalizability of research

- Referencing data in order to identify, discover and retrieve them

- Building and propagating knowledge

# ~~Why~~ Data Citation is Important ~~?~~

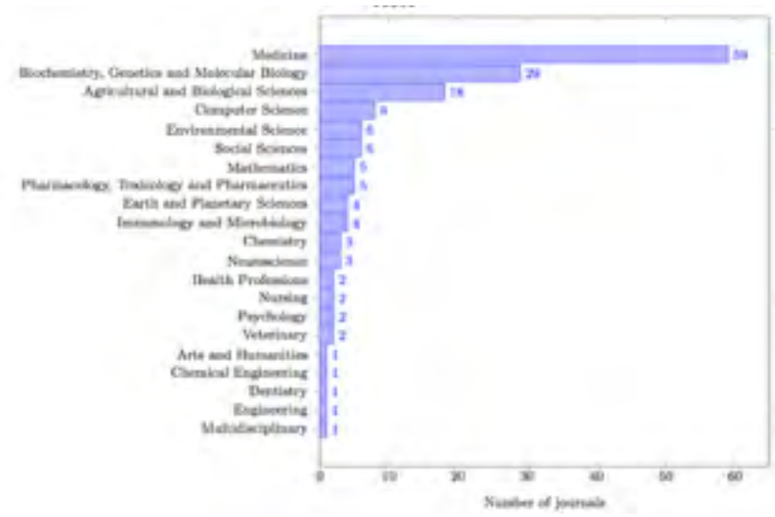
A lot of work has been done…

- Principles of data citation

- Recommendations for data citation systems

- Data publishing infrastructures and data journals

- Indexes and dataset impact

# ~~Why~~ Data Citation is Important ~~?~~

The practice of citing data is still not pervasive in scientific publishing

(euphemism)

**We need tools!**

**Computational Challenges in Data Citation**

*University of Pennsylvania, April 17-18, 2014*

*Workshop Report*

**Peter Buneman, Sarah Cohen-Boulakia, Susan B. Davidson, Jim Frew, Val Tannen**

## Introduction

Citation is an essential part of scientific publishing and, more generally, of scholarship. It is used to gauge the trust placed in published information and, for better or for worse, is an important factor in judging academic reputation. Now that so much scientific publishing involves data and takes place through a database rather than conventional journals, how is some part of a database to be cited? More generally, when one extracts some data from a large, complex, evolving database, how does one create the appropriate citation? How does one verify that the citation is correct?

Frameworks have been put forward by Information Scientists to serve as models or templates for citation. At the same time Data Scientists associated with various disciplines such as Bioinformatics, Earth Sciences, Neuroscience, etc., encounter interesting problems in trying to foster the citation of data. However, it is clear that for large evolving datasets and databases we are going to need algorithmic techniques and software technologies both to generate and to verify the correctness of citations, and these may well pose new problems for Computer Scientists.

The purpose of this workshop was to bring together people representing these different disciplines and enumerate the computational challenges and opportunities associated with data citation. The workshop was organized around three sessions – Citation Principles and Standards, Citation and Linked Open Data, and Executable Papers and Reproducibility – during which an overview talk was given followed by perspectives by participants. Participants then broke out into breakout groups, each of which contained people from different disciplines, and brainstormed what they believed to be the most important computational challenges for data citation. During a plenary session the next day, the challenges were revisited and refined. This report represents these findings.

In the remainder of this report, we discuss what data citations are and how they differ from citations to printed material as well as links. We then present the key computational challenges

# From the users perspective

# From the users perspective

- The generation of human- and machine-readable citations should be automatic

- Cited data should be uniquely identified: DOI will save the world

- Citing data should be easy: click, generate, copy and paste

- Setting up and maintaining a citation system should require low (no) effort to data creators/curators

# From the computer scientists perspective

# From the computer scientists perspective

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

- Data is not (always) fixed, it changes
- Persistent identifiers are (only) part of the solution
- Variable granularity (deep citations)
- Automatic generation of citations (yes, but how?)
- Different data types and formats

# This talk

- Focus: Automatic generation of human- and machine-readable citations

- Goal: To minimize the effort required to data creators and curators to setup and use a citation system

- What: A citation system for hierarchical data (XML)

# Use case: Digital archives

# What is an Archive?

# What is an Archive?

Shelves

# What is an Archive?

Shelves

Folders

# What is an Archive?

Shelves

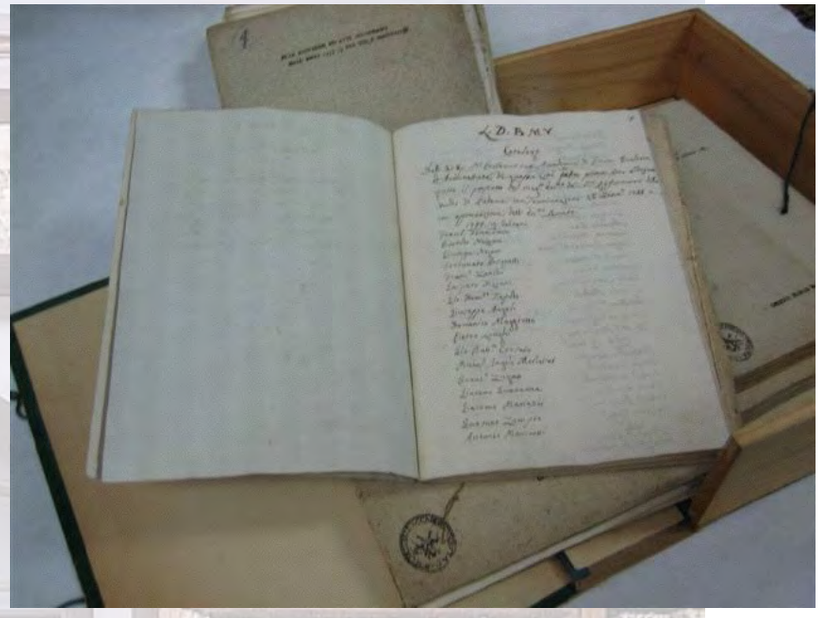
Envelops

# What is an Archive?

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Shelves

Folders

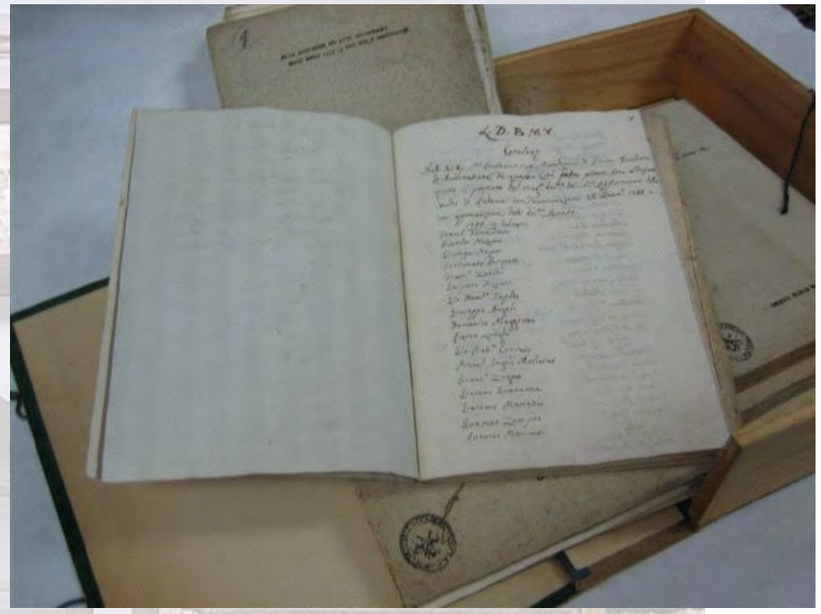


Envelops

# What is an Archive?

Shelves

Folders

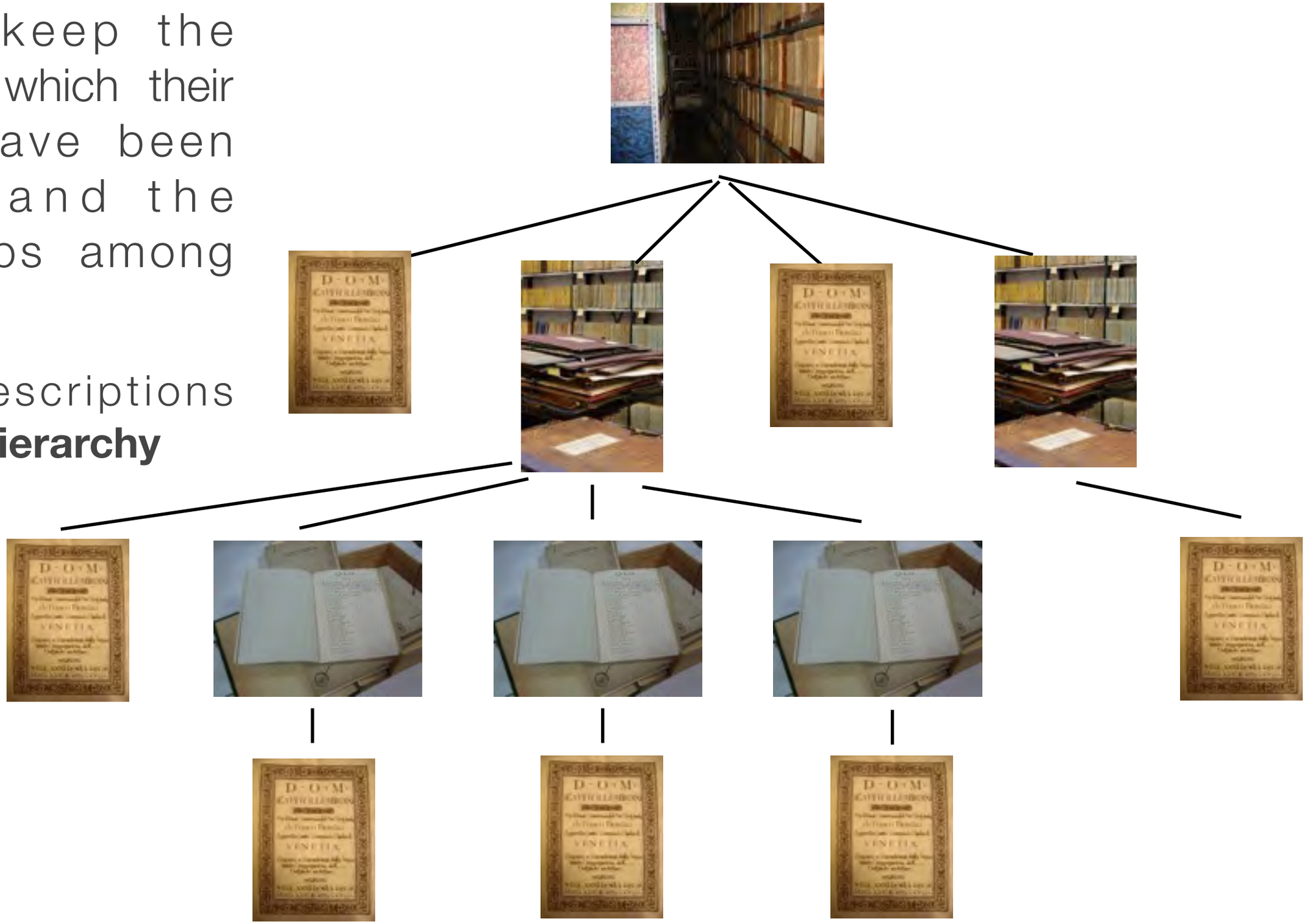


Envelops


Documents (e.g. letters, registers, testaments)

# Archival Tree

- Archives keep the **context** in which their records have been created and the relationships among them
- Archival descriptions constitute a **hierarchy**

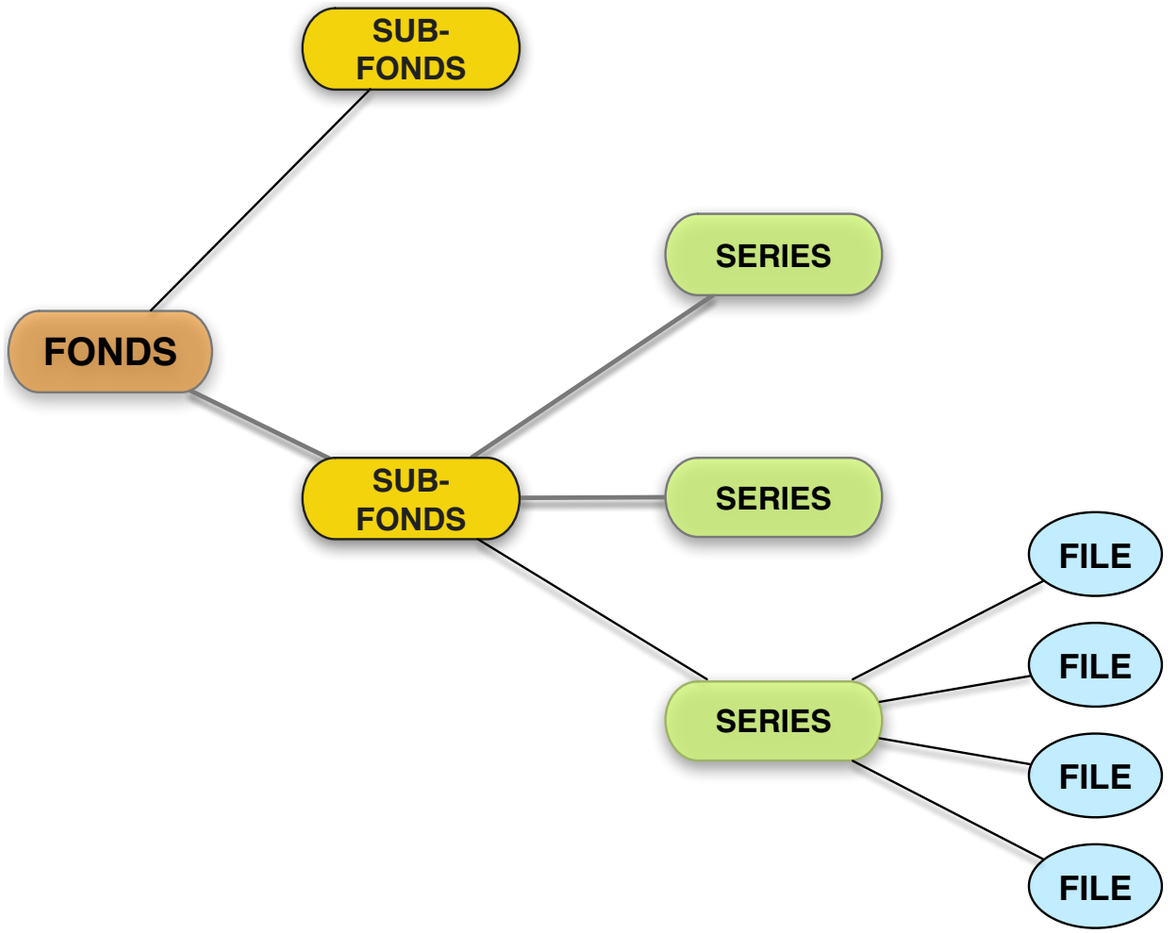


# Archival Tree

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

fonds
sub-fonds
series
document

# Encoding of archival data: EAD




(a) Archival Tree

```
<ead>
    <eadheader>
        […]
    </eadheader>
    <archdesc level="fonds">
        […]
        <did>[…]</did>
            <dsc level="fonds">
                […]
                <c01 level="sub-fonds">
                    […]
                </c01>
                <c01 level="sub-fonds">
                    […]
                    <c02 level="series">
                        […]
                    </c02>
                    <c02 level="series">
                        […]
                    </c02>
                    <c02 level="series">
                        […]
                        <c03 level="file">
                            […]
                        </c03>
                        <c03 level="file">
                            […]
                        </c03>
                        <c03 level="file">
                            […]
                        </c03>
                        <c03 level="file">
                            […]
                        </c03>
                    </c02>
                </c01>
            </dsc>
    </archdesc>
</ead>
```

(b) EAD representation

# Characteristics of EAD files

- A sigle EAD file encodes a whole archive

- “Big” XML files with deep hierarchy

- Heterogeneous use of tags across collections and within the same collection

- Every element and attribute of an EAD file is a potential citable unit

# EAD: Some statistics

| Collection | Files | Nodes | | Depth | | Size (KB) | | Max Fan Out | |
|---|---|---|---|---|---|---|---|---|---|
| | | max | median | max | median | max | median | max | median |
| AH 2005 | 233 | 14,648 | 158 | 21 | 6 | 760 | 15 | 1,332 | 23 |
| IISG 2005 | 798 | 52,213 | 513 | 17 | 9 | 2,290 | 34 | 2,601 | 21 |
| NA 2008 | 1681 | 160,061 | 880.5 | 18 | 9 | 9,750 | 58 | 10,271 | 34 |
| LoC 2014 | 2083 | 188,862 | 685 | 18 | 10 | 15,510 | 58 | 5,000 | 32 |
| UniMa 2014 | 662 | 69,766 | 711 | 10 | 8 | 2,960 | 40 | 6,861 | 43 |

AH 2005: UK Archival Hub, 2005 snapshot
IISG 2005: International Institute of Social History, 2005 snapshot
NA 2008: Nationaal Archief, The Netherlands, 2008 snapshot
Loc 2014: Library of Congress, 2014 snapshot
UniMa 2014: University of Maryland, 2014snapshot

# A Human-readable citation

Correspondence, 1951-1956,

"The Elements of Legal Theory" (unpublished). Books, box 135. Part II:

Writings (1905-1984), box 129-152. Huntington Cairns Papers.

Manuscript Division, Library of Congress.

http://hdl.loc.gov/loc.mss/eadmss.ms001024

# A Human-readable citation

Citable unit

```
Correspondence, 1951-1956,
```

Contextual Information (from ancestors of the citable unit)

```
"The Elements of Legal Theory" (unpublished). Books, box 135. Part II:

Writings (1905-1984), box 129-152. Huntington Cairns Papers.

Manuscript Division, Library of Congress.
```

```
http://hdl.loc.gov/loc.mss/eadmss.ms001024
```

(Persistent) Unique identifier of the EAD file

All the elements of the citations are obtained from the EAD file containing the citable unit

In general, EAD files always contain all the information required to build a citation and a citable unit alone cannot be used to create a complete citation

# A machine-readable citation

Conjunction of XPaths

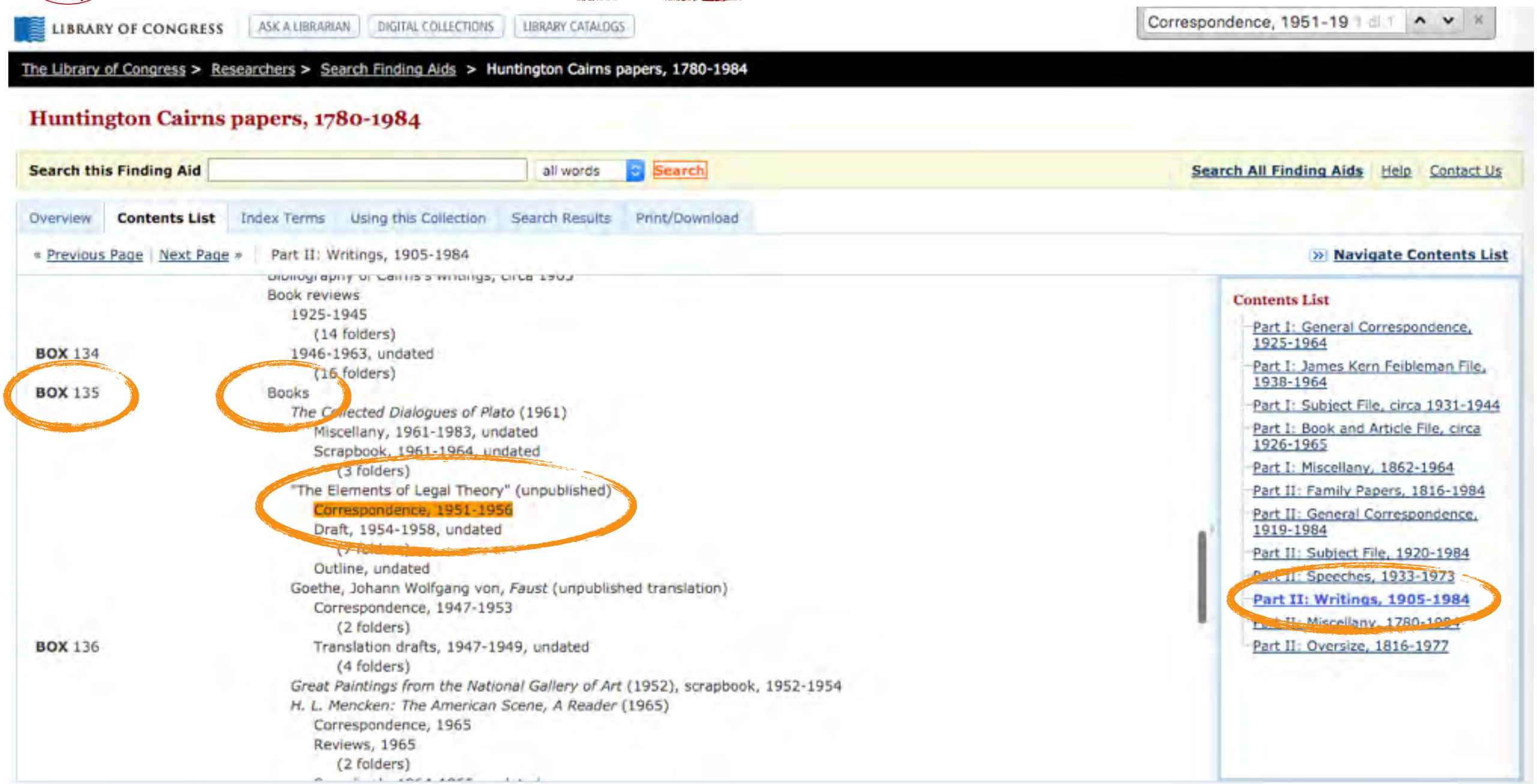
```
/ead/eadheader/eadid && /ead/eadheader/filedesc/publicationstmt/publisher && /ead/
archdesc/did/unittitle && /ead/archdesc/dsc/c01[10]/did/unittitle && /ead/archdesc/
dsc/c01[10]/did/unittitle/unitdate && /ead/archdesc/dsc/c01[10]/did/container/@type
&& /ead/archdesc/dsc/c01[10]/did/container &&  /ead/archdesc/dsc/c01[10]/c02/did/
container/@type && /ead/archdesc/dsc/c01[10]/c02/did/container && /ead/archdesc/dsc/
c01[10]/c02/did/unittitle &&  /ead/archdesc/dsc/c01[10]/c02/c03[4]/did/unittitle
&& /ead/archdesc/dsc/c01[10]/c02/c03[4]/did/container/@type && /ead/archdesc/dsc/
c01[10]/c02/c03[4]/did/container &&  /ead/archdesc/dsc/c01[10]/c02/c03[4]/c04[2]/
did/unittitle && /ead/archdesc/dsc/c01[10]/c02/c03[4]/c04[2]/c05[1]/did/unittitle
```

# A machine-readable citation

| Human-Readable Citation | Machine-Readable Citation |
|---|---|
| http://hdl.loc.gov/loc.mss/eadmss.ms001024 | /ead/eadheader/eadid |
| Manuscript Division, Library of Congress | /ead/eadheader/filedesc/publicationstmt/publisher |
| Huntington Cairns Papers | /ead/archdesc/did/unittitle |
| Part II: Writings | /ead/archdesc/dsc/c01[10]/did/unittitle |
| 1905-1984 | /ead/archdesc/dsc/c01[10]/did/unittitle/unitdate |
| box | /ead/archdesc/dsc/c01[10]/did/container/@type |
| 129-152 | /ead/archdesc/dsc/c01[10]/did/container |
| By Cairns | /ead/archdesc/dsc/c01[10]/c02[1]/did/unittitle |
| box | /ead/archdesc/dsc/c01[10]/c02[1]/did/container/@type |
| 129 | /ead/archdesc/dsc/c01[10]/c02[1]/did/container/ |
| Books | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/unittitle |
| box | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container/@type |
| 135 | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container |
| "The Elements of Legal Theory" (unpublished) | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/did/unittitle |
| Correspondence, 1951-1956 | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/c05[1]/did/unittitle |

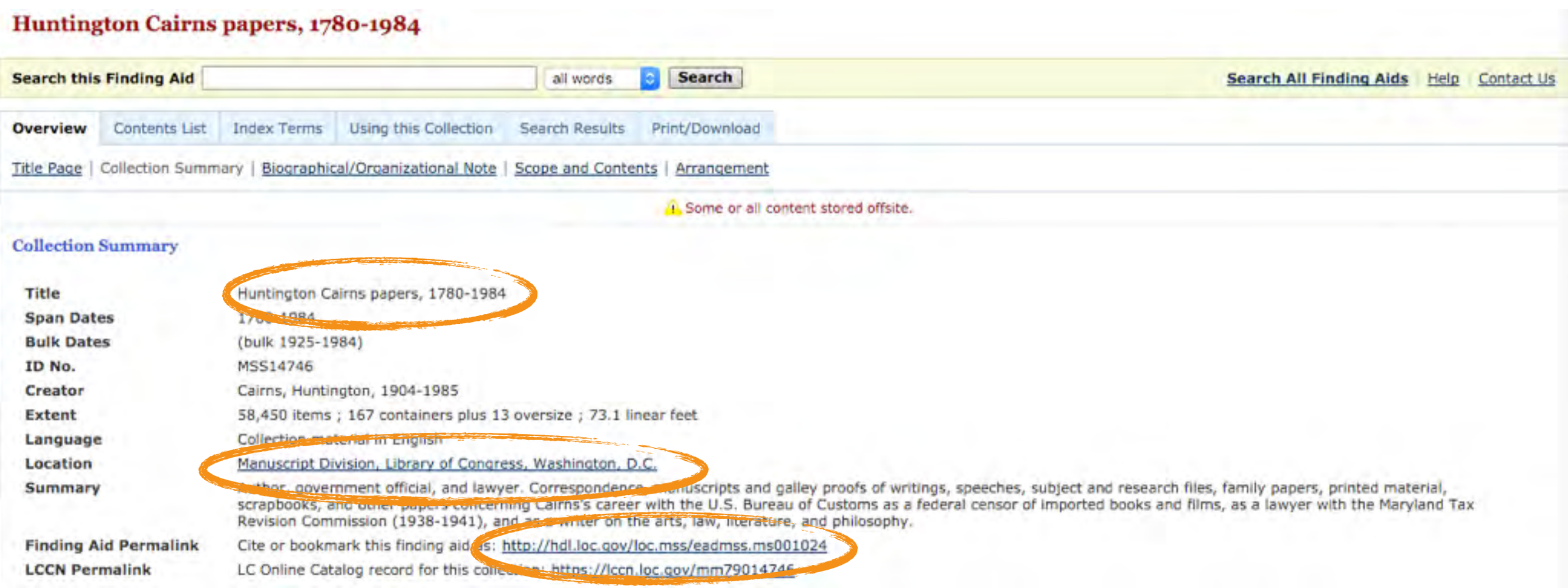
# What does the user see?






Correspondence, 1951-1956,
"The Elements of Legal Theory" (unpublished). Books, box 135. Part II: Writings (1905-1984), box 129-152. Huntington Cairns Papers.
Manuscript Division, Library of Congress.
http://hdl.loc.gov/loc.mss/eadmss.ms001024

# What does the user see?



**Huntington Cairns papers, 1780-1984**

Search this Finding Aid | all words | Search | Search All Finding Aids | Help | Contact Us

Overview | Contents List | Index Terms | Using this Collection | Search Results | Print/Download

Title Page | Collection Summary | Biographical/Organizational Note | Scope and Contents | Arrangement

Some or all content stored offsite.

**Collection Summary**

| | |
|---|---|
| **Title** | Huntington Cairns papers, 1780-1984 |
| **Span Dates** | 1780-1984 |
| **Bulk Dates** | (bulk 1925-1984) |
| **ID No.** | MSS14746 |
| **Creator** | Cairns, Huntington, 1904-1985 |
| **Extent** | 58,450 items ; 167 containers plus 13 oversize ; 73.1 linear feet |
| **Language** | Collection material in English |
| **Location** | Manuscript Division, Library of Congress, Washington, D.C. |
| **Summary** | Author, government official, and lawyer. Correspondence, manuscripts and galley proofs of writings, speeches, subject and research files, family papers, printed material, scrapbooks, and other papers concerning Cairns's career with the U.S. Bureau of Customs as a federal censor of imported books and films, as a lawyer with the Maryland Tax Revision Commission (1938-1941), and as a writer on the arts, law, literature, and philosophy. |
| **Finding Aid Permalink** | Cite or bookmark this finding aid as: http://hdl.loc.gov/loc.mss/eadmss.ms001024 |
| **LCCN Permalink** | LC Online Catalog record for this collection: https://lccn.loc.gov/mm79014746 |

```
Correspondence, 1951-1956,
"The Elements of Legal Theory" (unpublished). Books, box 135. Part II: Writings
(1905-1984), box 129-152. Huntington Cairns Papers.
Manuscript Division, Library of Congress.
http://hdl.loc.gov/loc.mss/eadmss.ms001024
```

# Generation of citations: The problem

Given:

**data**

```
/ead/archdesc/dsc/c01[10]/
c02/c03[4]/c04[2]/c05[1]/
did/unittitle
```

Query: XPath


Dataset: EAD file

generate

**citation**

| Human-Readable Citation | Machine-Readable Citation |
|---|---|
| http://hdl.loc.gov/loc.mss/eadmss.ms001024 | /ead/eadheader/eadid |
| Manuscript Division, Library of Congress | /ead/eadheader/filedesc/publicationstmt/publisher |
| Huntington Cairns Papers | /ead/archdesc/did/unittitle |
| Part II: Writings | /ead/archdesc/dsc/c01[10]/did/unittitle |
| 1905-1984 | /ead/archdesc/dsc/c01[10]/did/unittitle/unitdate |
| box | /ead/archdesc/dsc/c01[10]/did/container/@type |
| 129-152 | /ead/archdesc/dsc/c01[10]/did/container |
| By Cairns | /ead/archdesc/dsc/c01[10]/c02[1]/did/unittitle |
| box | /ead/archdesc/dsc/c01[10]/c02[1]/did/container/@type |
| 129 | /ead/archdesc/dsc/c01[10]/c02[1]/did/container/ |
| Books | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/unittitle |
| box | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container/@type |
| 135 | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container |
| "The Elements of Legal Theory" (unpublished) | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/did/unittitle |
| Correspondence, 1951-1956 | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/c05[1]/did/unittitle |

# Learning to cite framework

# Learning to cite framework

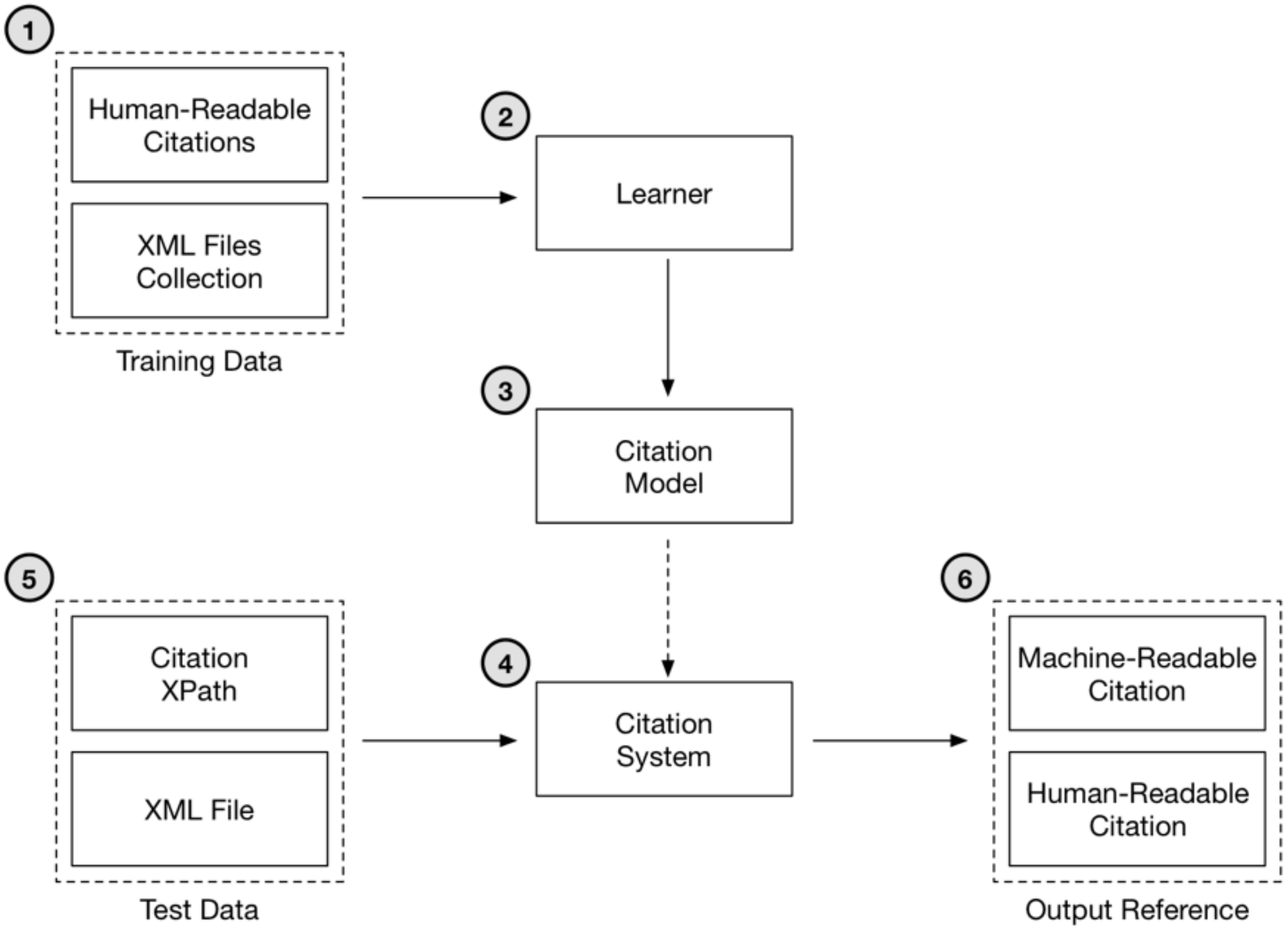
- The idea is to employ a machine learning approach for the generation of citations

- Learn from some sample data (human-readable citations), get a citation model out of it, and generate citations

- Require low effort (and resources) to data creators and curators

- Handle data heterogeneity

# Learning to cite framework



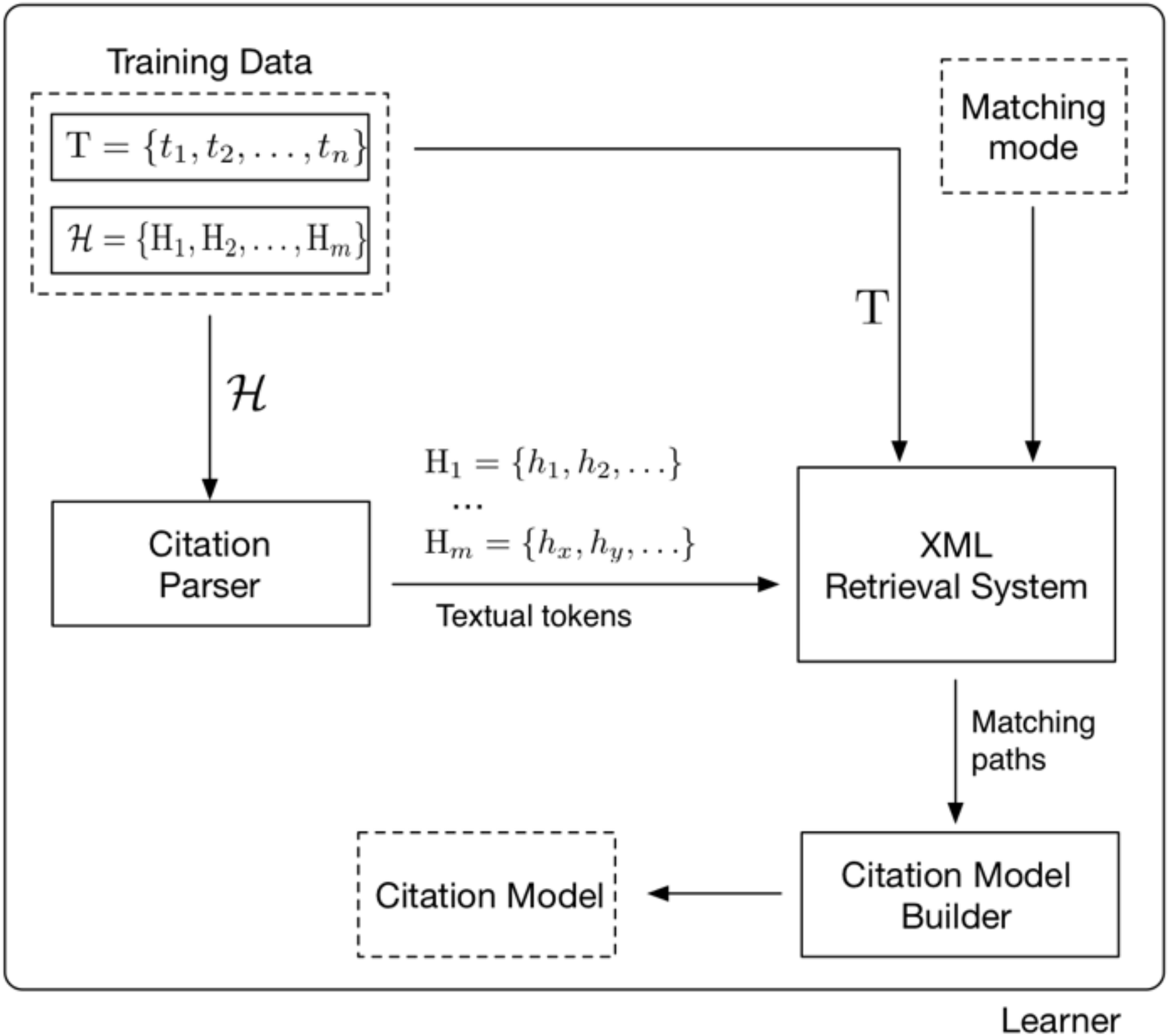
# Learning to cite (LtC) framework

- Two phases: Training and Validation

- Training phase: Learn the citation model from the training data

- Validation phase: Optimization of model parameters according to an evaluation measure

# LtC: Learner


Training Data
$\mathrm{T} = \{t_1, t_2, \ldots, t_n\}$
$\mathcal{H} = \{\mathrm{H}_1, \mathrm{H}_2, \ldots, \mathrm{H}_m\}$
$\mathcal{H}$
Citation Parser
$\mathrm{H}_1 = \{h_1, h_2, \ldots\}$
...
$\mathrm{H}_m = \{h_x, h_y, \ldots\}$
Textual tokens
T
Matching mode
XML Retrieval System
Matching paths
Citation Model Builder
Citation Model
Learner

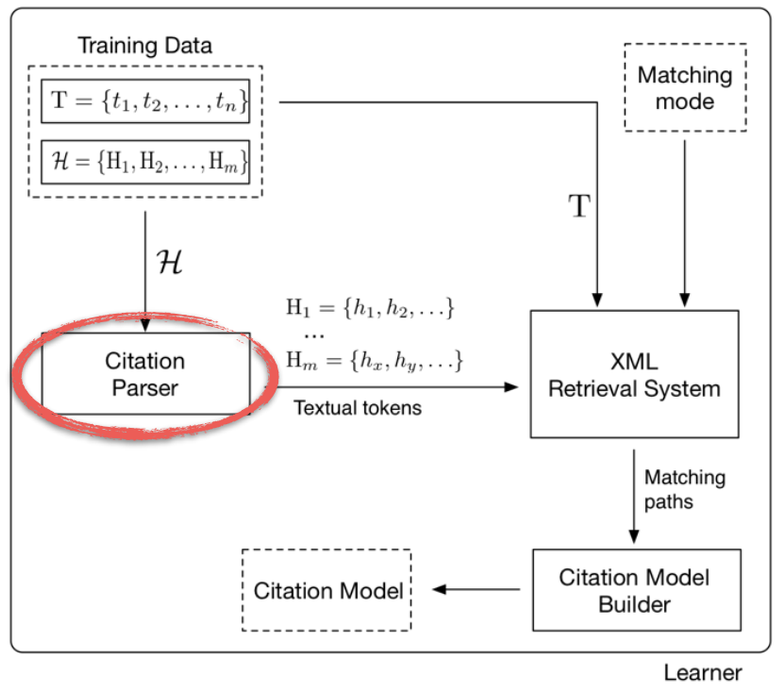
# Learner: citation parser

```
Correspondence, 1951-1956,
"The Elements of Legal Theory" (unpublished). Books, box 135. Part II: Writings
(1905-1984), box 129-152. Huntington Cairns Papers.
Manuscript Division, Library of Congress.
http://hdl.loc.gov/loc.mss/eadmss.ms001024
```

parser

$$H_j \in \mathcal{H} \xrightarrow{\text{parser}} H_j = \{h_1, h_2, \ldots, h_n\}$$

```
Correspondence, 1951-1956
"The Elements of Legal Theory" (unpublished)
Books
box
135
Part II: Writings
1905-1984
box
129-152
Huntington Cairns Papers
Manuscript Division, Library of Congress
http://hdl.loc.gov/loc.mss/eadmss.ms001024
```



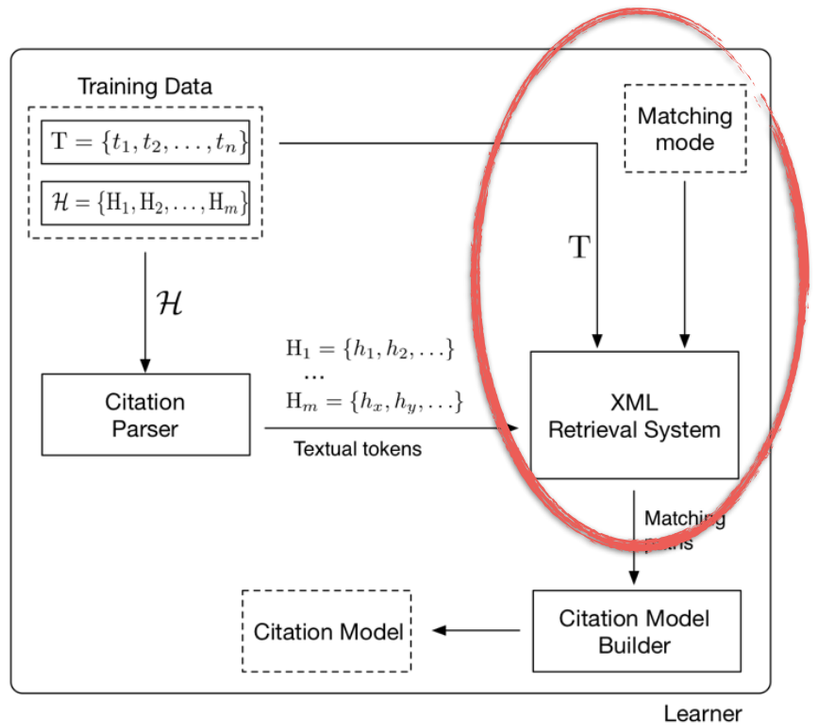
# Learner: XML retrieval system

```
Correspondence, 1951-1956
"The Elements of Legal Theory" (unpublished)
Books
box
135
Part II: Writings
1905-1984
box
129-152
Huntington Cairns Papers
Manuscript Division, Library of Congress
http://hdl.loc.gov/loc.mss/eadmss.ms001024
```

Retrieval →


EAD file

- Seeks the textual tokens in the XML file
- Returns the XPath of the matching elements + similarity scores



# Learner: XML retrieval system

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

## Textual tokens

```
Correspondence, 1951-1956
"The Elements of Legal
Theory" (unpublished)
Books
box
135
Part II: Writings
1905-1984
box
129-152
Huntington Cairns Papers
Manuscript Division, Library of
Congress
http://hdl.loc.gov/loc.mss/
eadmss.ms001024
```

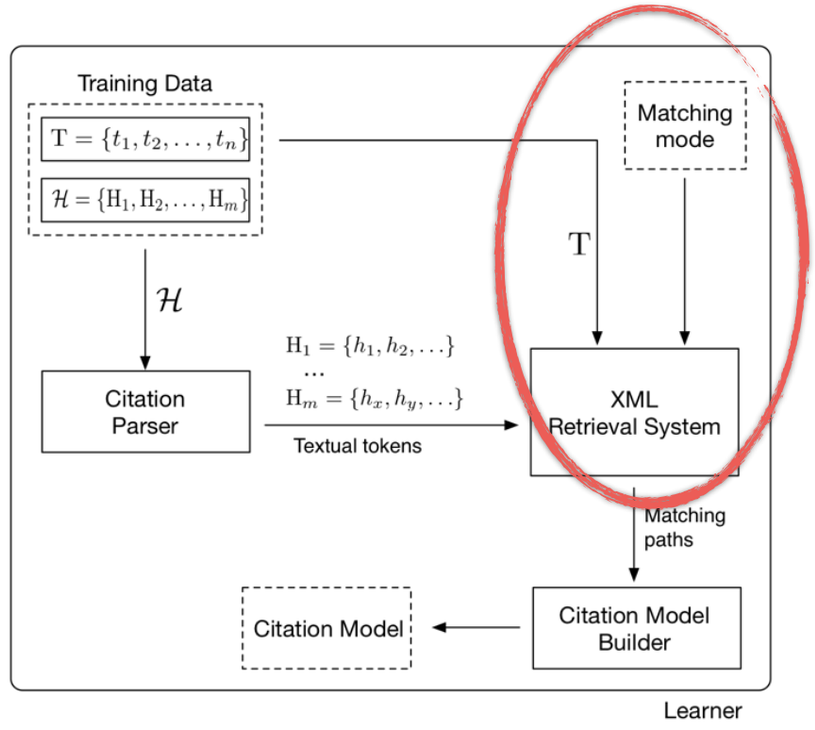
Match

## EAD file

# Learner: XML retrieval system

- Matching mode:
  - **Exact match mode**: retrieves those elements containing all and only the words in the given token
  - **Shallow match mode**: retrieves those elements containing all but not only the words in the given token;
  - **Mixed match mode**: uses the exact match mode first and if no result is returned it uses the shallow mode

The aim of shallow match modes* is to retrieve more candidate XPaths to give more flexibility to the citation model



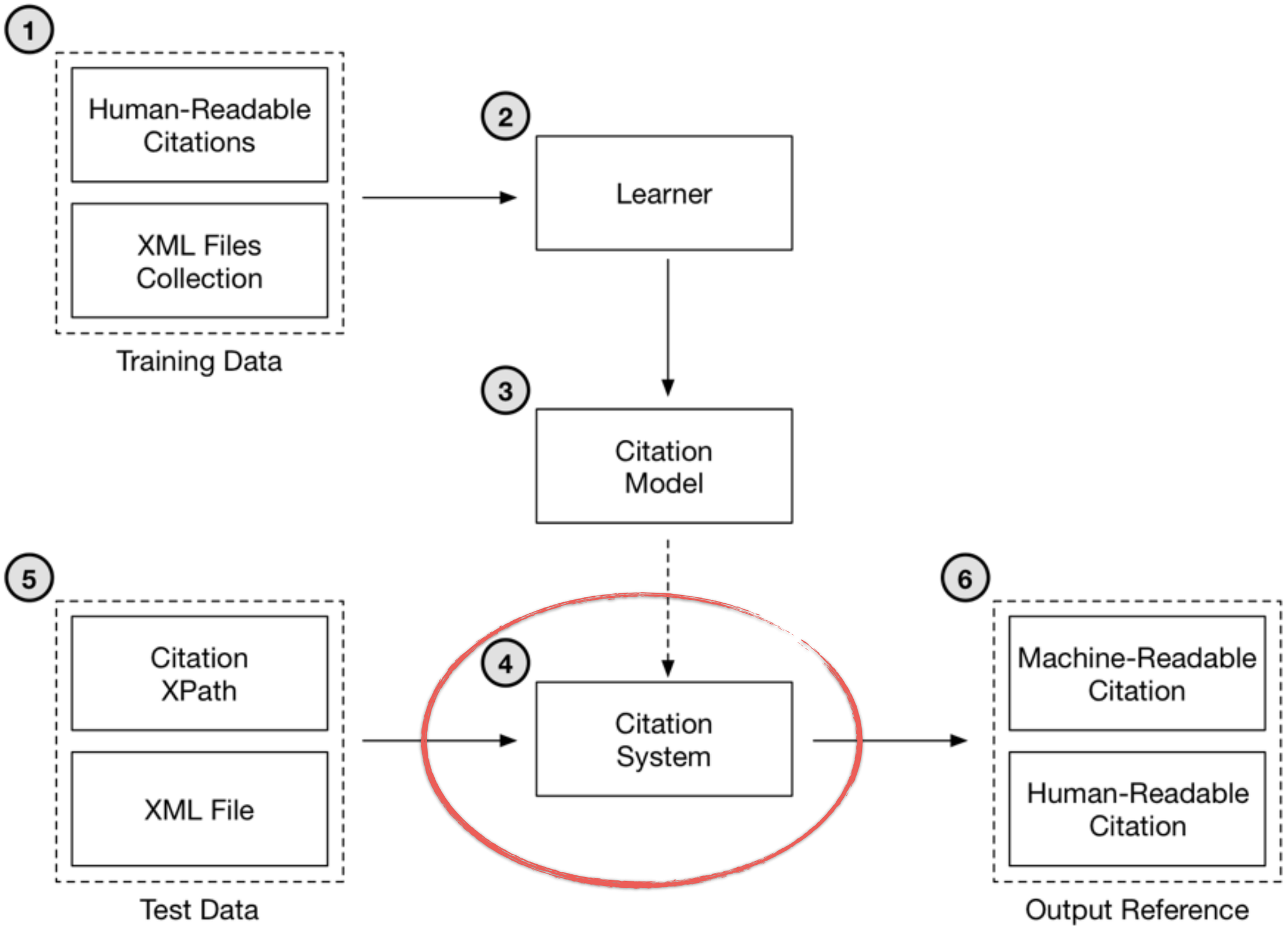

*There are other (shallower) matching modes

# Learner: Citation Model

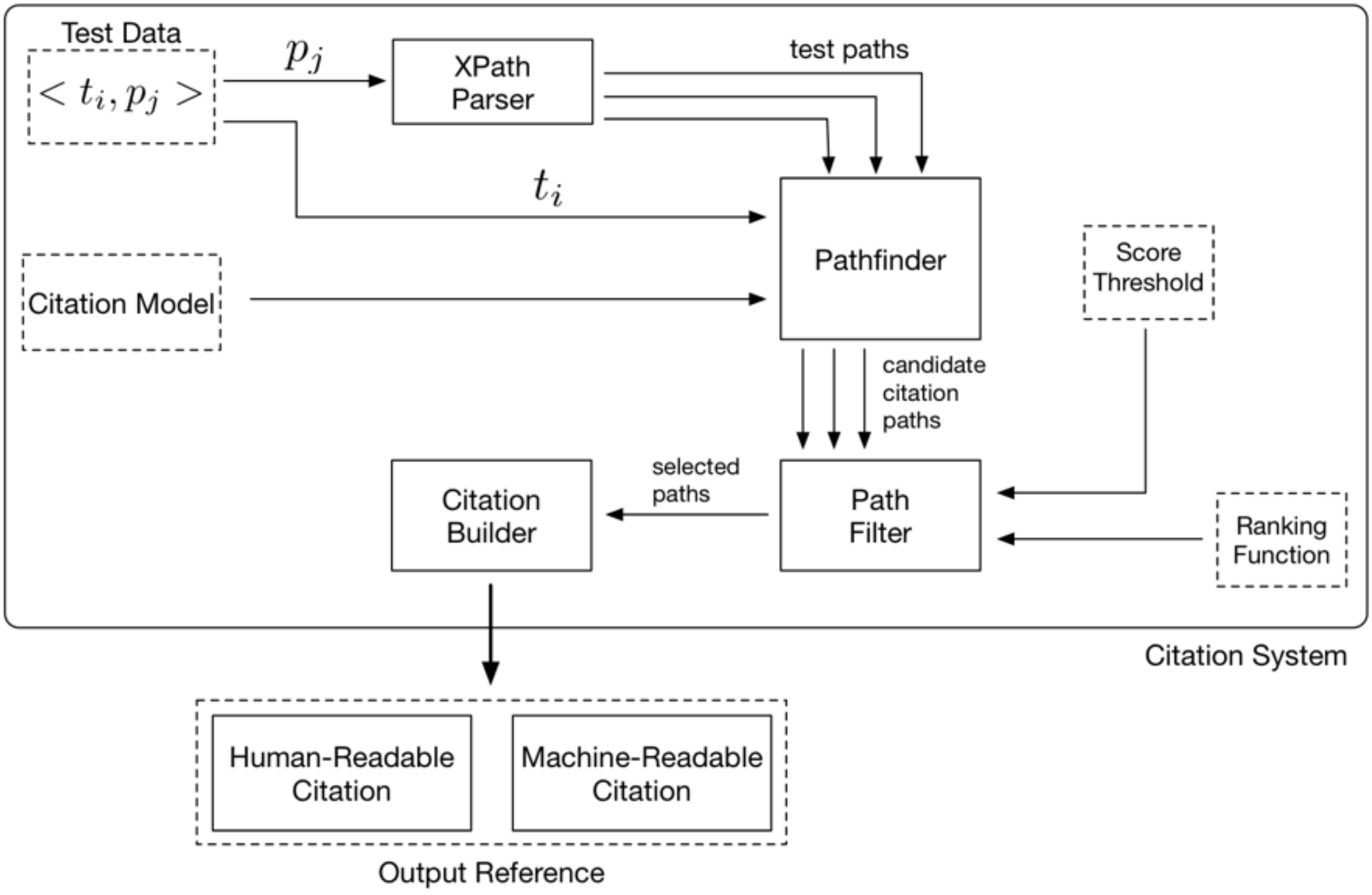
```
<cite-tree>
    <ead score="0" frequency="0" type="element">
        <eadheader score="0" frequency="0" type="element">
            <eadid score="1.0" frequency="1" type="element"/>
            <filedesc score="0" frequency="0" type="element">
                <publicationstmt score="0" frequency="0" type="element">
                    <publisher score="0.9413807890996732" frequency="1.0" type="element">
                        <extptr score="0.9413807890996732" frequency="1" type="element"/>
                    </publisher>
                </publicationstmt>
                <titlestmt score="0" frequency="0" type="element">
                    <titleproper score="0.9311609000928631" frequency="1" type="element"/>
                </titlestmt>
            </filedesc>
        </eadheader>
        <archdesc score="0" frequency="0" type="element">
            <did score="0" frequency="0" type="element">
                <unittitle score="0.9740960467331898" frequency="1" type="element">
                    <unitdate score="0.9740960467331898" frequency="1" type="element"/>
                </unittitle>
            </did>
            <dsc score="0" frequency="0" type="element">
                <c01 score="0" frequency="0" type="element">
                    <did score="0" frequency="0" type="element">
                        <unittitle score="0.9197168697545395" frequency="1" type="element">
                            <unitdate score="0.9369796895969322" frequency="2.0" type="element"/>
                        </unittitle>
                        <container score="0.9357849740192015" frequency="1.0" type="element">
                            <type score="1.0" frequency="3.0" type="attribute"/>
                        </container>
                    </did>
                    <c02 score="0" frequency="0" type="element">
                        <did score="0" frequency="0" type="element">
                            <container score="1.0" frequency="1.0" type="element">
                                <type score="1.0" frequency="3.0" type="attribute"/>
                            </container>
                            <unittitle score="0.9542425094393249" frequency="1" type="element"/>
                        </did>
                        […]
                    </c02>
                </c01>
            </dsc>
        </archdesc>
    </ead>
</cite-tree>
```

sample citation model created from a single human-readable citation

# Learning to cite framework



# Citation system



# System: Test data and parser

Sample test XPath identifying the citable unit

```
/ead/archdesc/dsc/c01[10]/did/unittitle
```

- The idea is that every element in the input file identified by any possible sub-path (i.e., for each location step) and its descendants may contain useful information to build the citation

```
(1)/ead/archdesc/dsc/c01/did/unittitle
(2)/ead/archdesc/dsc/c01/did
(3) /ead/archdesc/dsc/c01
(4)/ead/archdesc/dsc
(5)/ead/archdesc
(6)/ead
```

- Test paths do not take into account specific indexes and predicates
- The problem is: How do we select only the relevant information from the elements identified by the sub-XPaths and their descendants?

# System: Pathfinder

- Each test XPath is matched with the citation model:
  - **Exact match**: the XPath (+score and frequency) of the element is returned along with all its descendants with score>0
    - /ead/archdesc/dsc/c01 returns 7 candidate paths
    - /ead returns 13 candidate paths
  - **Best match**: Given an XPath we seek the element identified by the deepest location step (`unittitle`); if there is a match, then we seek the longest path within the XPath with a match in the citation model; if there is more than one match, then only the longest path is kept.

# System: Pathfinder (example)

test path

/ead/archdesc/dsc/c01

exact match + descendants

```
<cite-tree>
    <ead score="0" frequency="0" type="element">
        <eadheader score="0" frequency="0" type="element">
            <eadid score="1.0" frequency="1" type="element"/>
            <filedesc score="0" frequency="0" type="element">
                <publicationstmt score="0" frequency="0" type="element">
                    <publisher score="0.9413807890996732" frequency="1.0" type="element">
                        <extptr score="0.9413807890996732" frequency="1" type="element"/>
                    </publisher>
                </publicationstmt>
                <titlestmt score="0" frequency="0" type="element">
                    <titleproper score="0.9311609000928631" frequency="1" type="element"/>
                </titlestmt>
            </filedesc>
        </eadheader>
        <archdesc score="0" frequency="0" type="element">
            <did score="0" frequency="0" type="element">
                <unittitle score="0.9740960467331898" frequency="1" type="element">
                    <unitdate score="0.9740960467331898" frequency="1" type="element"/>
                </unittitle>
            </did>
            <dsc score="0" frequency="0" type="element">
                <c01 score="0" frequency="0" type="element">
                    <did score="0" frequency="0" type="element">
                        <unittitle score="0.9197168697545395" frequency="1" type="element">
                            <unitdate score="0.9369796895969322" frequency="2.0" type="element"/>
                        </unittitle>
                        <container score="0.9357849740192015" frequency="1.0" type="element">
                            <type score="1.0" frequency="3.0" type="attribute"/>
                        </container>
                    </did>
                    <c02 score="0" frequency="0" type="element">
                        <did score="0" frequency="0" type="element">
                            <container score="1.0" frequency="1.0" type="element">
                                <type score="1.0" frequency="3.0" type="attribute"/>
                            </container>
                            <unittitle score="0.9542425094393249" frequency="1" type="element"/>
                        </did>
                        […]
                    </c02>
                </c01>
            </dsc>
        </archdesc>
    </ead>
</cite-tree>
```

# System: Pathfinder (example)

test path

/ead/archdesc/dsc/c01

exact match + descendants

candidate paths

/ead/archdesc/dsc/c01/did/unittitle
/ead/archdesc/dsc/c01/did/unittitle/unitdate
/ead/archdesc/dsc/c01/did/container
/ead/archdesc/dsc/c01/did/container/@type
/ead/archdesc/dsc/c01/c02/did/container
/ead/archdesc/dsc/c01/c02/did/container/@type
/ead/archdesc/dsc/c01/c02/did/unittitle

```
<unittitle score="0.9197168697545395" frequency="1" type="element">
    <unitdate score="0.9369796895969322" frequency="2.0" type="element"/>
</unittitle>
<container score="0.9357849740192015" frequency="1.0" type="element">
    <type score="1.0" frequency="3.0" type="attribute"/>
</container>
</did>
<c02 score="0" frequency="0" type="element">
    <did score="0" frequency="0" type="element">
        <container score="1.0" frequency="1.0" type="element">
            <type score="1.0" frequency="3.0" type="attribute"/>
        </container>
        <unittitle score="0.9542425094393249" frequency="1" type="element"/>
```

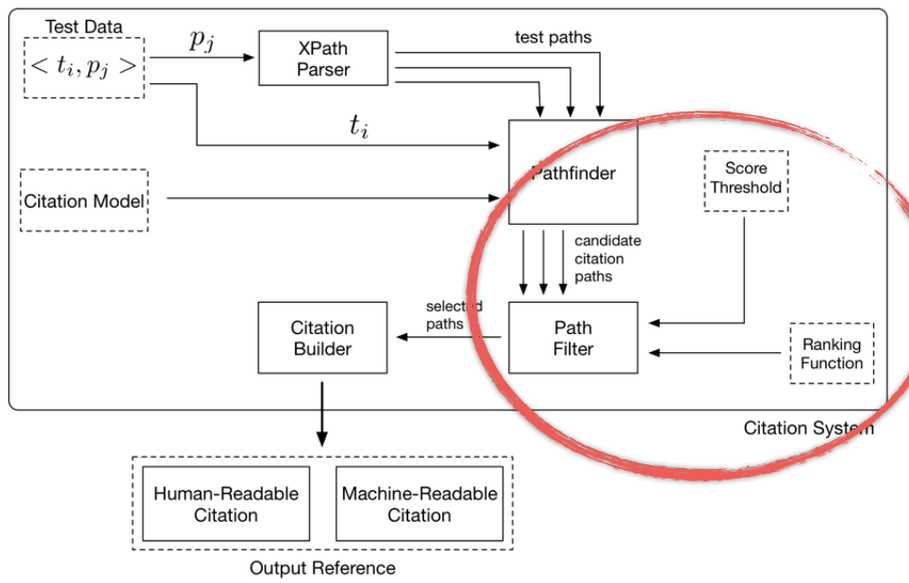
# System: Path filter

- Each candidate XPath comes with a **frequency**, a **score** and a **relative depth** (*relDepth*) which indicates the distance from the element identified by the candidate path and the element identified by the test path considered at the moment

test path (1): `/ead/archdesc/dsc/c01/`

candidate path: `/ead/archdesc/dsc/c01/unittitle/unitdate`

relDepth = 2



# System: Path filter

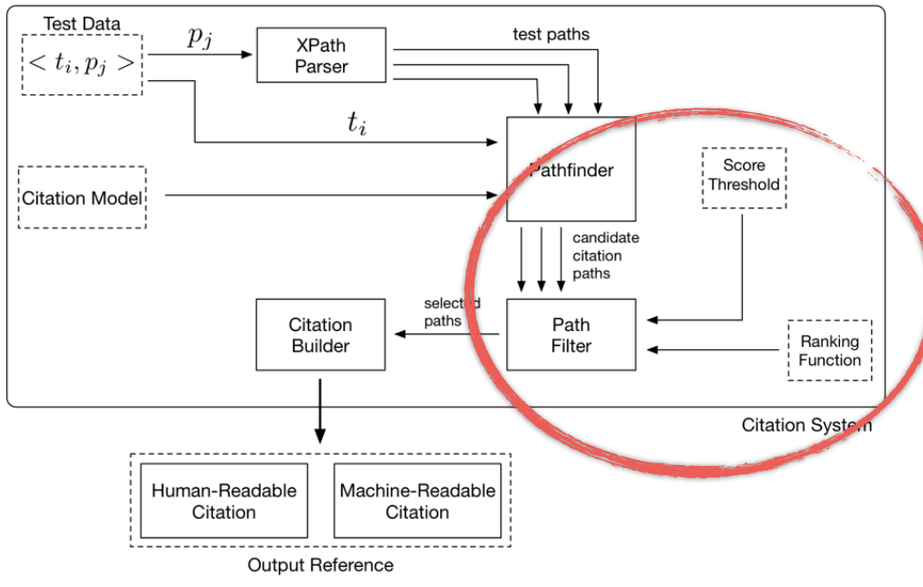
- Each candidate XPath comes with a **frequency**, a **score** and a **relative depth** (*relDepth*) which indicates the dis from the element identified by the candidate path e element identified by the test path consider the moment

test path (1): `/ead/archdesc/dsc/c01/`

candidate path: `/ead/archdesc/ds ittitle/unitdate`

We need to rank the candidate paths and keep only the most relevant ones

relDepth = 2



# System: Path filter (ranking function)

Frequency Score Depth Normalization (FSDN): $\frac{\text{score} * \text{frequency}}{\text{relDepth}}$

Score Depth Normalization (SDN): $\frac{\text{score}}{\text{relDepth}}$

Frequency Depth Normalization (FDN): $\frac{\text{frequency}}{\text{relDepth}}$

Frequency Score (FS): $\text{score} * \text{frequency}$

The scores are further normalized in [0,1] and only those above a given threshold are used to build the final citation

# System: Path filter (example)

| Candidate paths | FDSN score |
|---|---|
| /ead/archdesc/dsc/c01/did/unittitle | 0.459 |
| /ead/archdesc/dsc/c01/did/unittitle/unitdate | 0.625 |
| /ead/archdesc/dsc/c01/did/container | 0.468 |
| /ead/archdesc/dsc/c01/did/container/@type | 1.000 |
| /ead/archdesc/dsc/c01/c02/did/container | 0.333 |
| /ead/archdesc/dsc/c01/c02/did/container/@type | 0.750 |
| /ead/archdesc/dsc/c01/c02/did/unittitle | 0.310 |

# System: Path filter (example)

| Candidate paths | FDSN score | |
|---|---|---|
| /ead/archdesc/dsc/c01/did/container/@type | 1.000 | |
| /ead/archdesc/dsc/c01/c02/did/container/@type | 0.750 | |
| /ead/archdesc/dsc/c01/did/unittitle/unitdate | 0.625 | |
| /ead/archdesc/dsc/c01/did/container | 0.468 | |
| /ead/archdesc/dsc/c01/did/unittitle | 0.459 | score threshold = 0.450 |
| /ead/archdesc/dsc/c01/c02/did/container | 0.333 | |
| /ead/archdesc/dsc/c01/c02/did/unittitle | 0.310 | |

Finally, the selected candidate paths are enriched with the indexes and predicates from the original query

/ead/archdesc/dsc/c01[10]/did/container/@type
/ead/archdesc/dsc/c01[10]/c02/did/container/@type
/ead/archdesc/dsc/c01[10]/did/unittitle/unitdate
/ead/archdesc/dsc/c01[10]/did/container

work done by the **citation builder** component

# Validation phase

- It is required to optimize the model parameters: *matching mode*, *ranking function* and *score threshold*

- k-folds cross validation is used

- We define 3 optimization measures: *precision*, *recall* and *f-score*

$$\text{precision} = \frac{\text{correct paths returned}}{\text{total paths returned}} \quad \text{(correctness)}$$

$$\text{recall} = \frac{\text{correct paths returned}}{\text{total correct paths}} \quad \text{(completeness)}$$

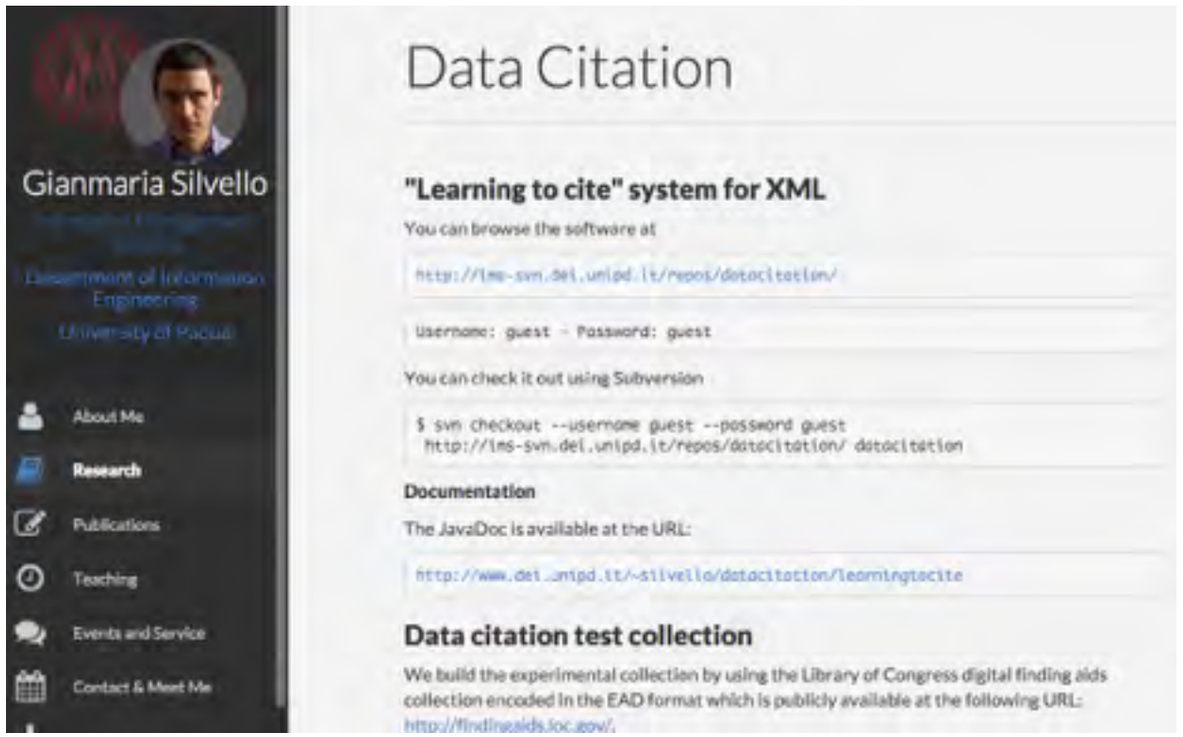
$$\text{fscore} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

# System implementation and data

- The citation system is open-source and implemented in Java (Maven project) as well as the code for the experiments

- The training data, test data and the ground truth are openly available

- http://www.dei.unipd.it/~silvello/datacitation/



# Experimental Evaluation

# Experimental data

- Based on the *Library of Congress EAD* collection (2083 files); several sub-collections; different archivists; 11M citable units (5k min - 385k max)

- *Training data*: 100 human-readable citations and EAD files

- *Validation data:* a subset of the training (5-folds validation)

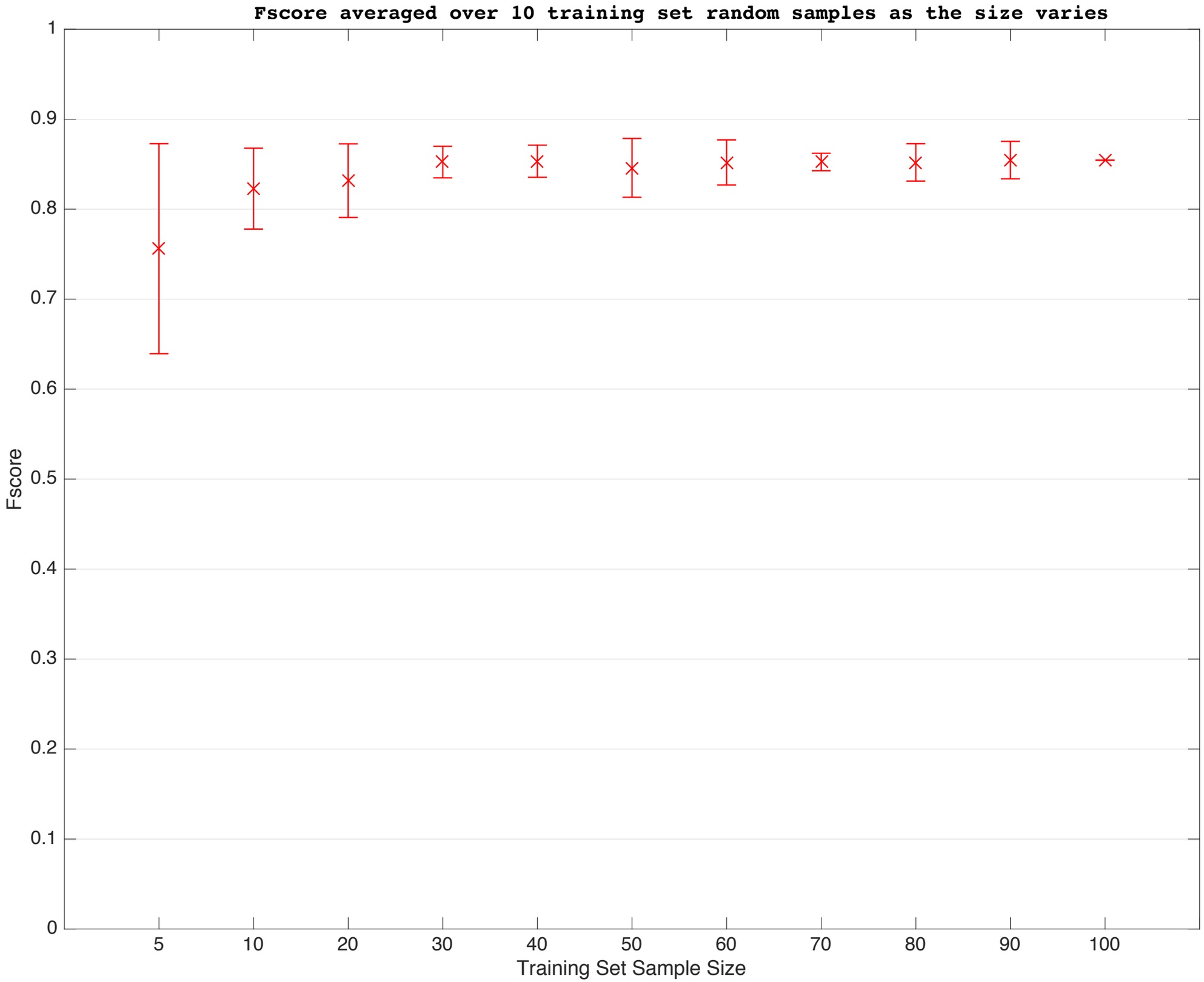
- *Test data:* 50 XPaths identifying citable units and EAD files (not in the training set)

- *Ground truth*: 150 “correct” human- and machine-readable citations

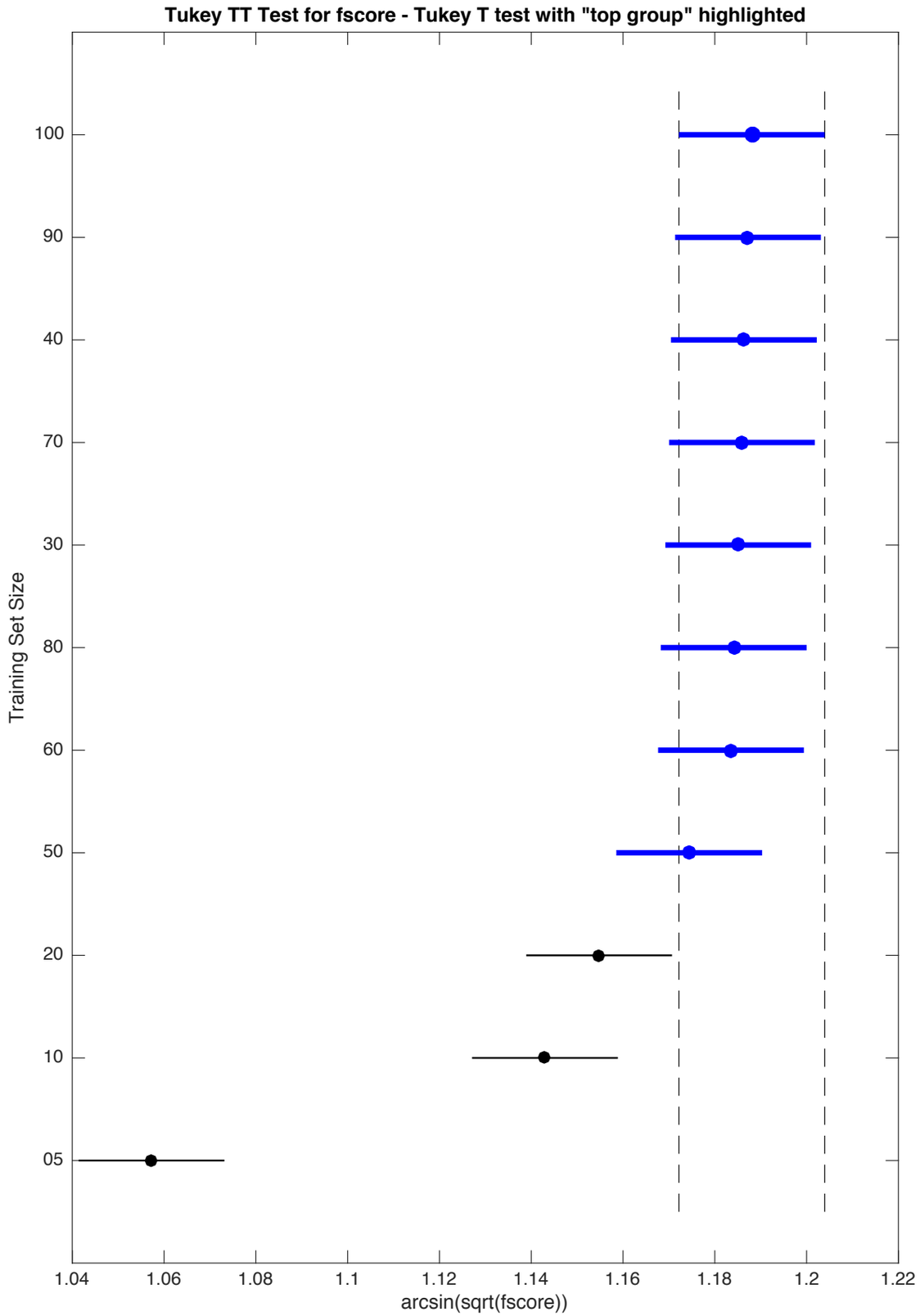
# Effect of parameters selection

| Tree Type | Ranking Function | Score Threshold | Avg Precision | Std Precision | Avg Recall | Std Recall | Avg Fscore | Std Fscore |
|---|---|---|---|---|---|---|---|---|
| **exact** | FDN | 0.1 | 0.3789 | 0.06 | 0.8975 | 0.04 | 0.5231 | 0.04 |
| **exact** | FDN | 0.5 | 0.7356 | 0.01 | 0.7448 | 0.03 | **0.7316** | 0.01 |
| **exact** | FDN | 1.0 | **0.7908** | 0.04 | 0.4552 | 0.05 | 0.5702 | 0.04 |
| **exact** | FS | 0.1 | 0.3813 | 0.07 | 0.8962 | 0.03 | 0.5196 | 0.04 |
| **exact** | FS | 0.5 | 0.6042 | 0.01 | 0.6919 | 0.03 | 0.6372 | 0.01 |
| **exact** | FS | 1.0 | 0.7211 | 0.02 | 0.2949 | 0.05 | 0.4087 | 0.03 |
| **exact** | FSDN | 0.1 | 0.3769 | 0.06 | 0.8975 | 0.04 | 0.5208 | 0.04 |
| **exact** | FSDN | 0.5 | 0.7293 | 0.01 | 0.7440 | 0.03 | 0.7278 | 0.01 |
| **exact** | FSDN | 1.0 | **0.7908** | 0.04 | 0.4542 | 0.08 | 0.5694 | 0.05 |
| **exact** | SDN | 0.1 | 0.1845 | 0.04 | **0.9052** | 0.04 | 0.3014 | 0.04 |
| **exact** | SDN | 0.5 | 0.2607 | 0.00 | 0.7684 | 0.04 | 0.3857 | 0.01 |
| **exact** | SDN | 1.0 | 0.3564 | 0.01 | 0.3411 | 0.04 | 0.3411 | 0.02 |
| **mixed** | FDN | 0.1 | 0.3186 | 0.05 | 0.8942 | 0.04 | 0.4631 | 0.04 |
| **mixed** | FDN | 0.5 | 0.5957 | 0.02 | 0.7111 | 0.05 | 0.6403 | 0.03 |
| **mixed** | FDN | 1.0 | 0.6115 | 0.04 | 0.3636 | 0.04 | 0.4477 | 0.03 |
| **mixed** | FS | 0.1 | 0.3339 | 0.08 | 0.8901 | 0.05 | 0.4734 | 0.06 |
| **mixed** | FS | 0.5 | 0.6127 | 0.03 | 0.6473 | 0.04 | 0.6220 | 0.03 |
| **mixed** | FS | 1.0 | 0.7028 | 0.04 | 0.2990 | 0.10 | 0.4095 | 0.06 |
| **mixed** | FSDN | 0.1 | 0.3276 | 0.05 | 0.8942 | 0.04 | 0.4718 | 0.04 |
| **mixed** | FSDN | 0.5 | 0.6514 | 0.02 | 0.7252 | 0.05 | 0.6789 | 0.03 |
| **mixed** | FSDN | 1.0 | 0.7746 | 0.03 | 0.4472 | 0.05 | 0.5581 | 0.04 |
| **mixed** | SDN | 0.1 | 0.1469 | 0.05 | **0.9045** | 0.04 | 0.2493 | 0.05 |
| **mixed** | SDN | 0.5 | 0.2690 | 0.01 | 0.7676 | 0.05 | 0.3948 | 0.01 |
| **mixed** | SDN | 1.0 | 0.4234 | 0.01 | 0.3643 | 0.05 | 0.3822 | 0.02 |
| **shallow** | FDN | 0.1 | 0.1630 | 0.04 | 0.8679 | 0.04 | 0.2719 | 0.04 |
| **shallow** | FDN | 0.5 | 0.3645 | 0.02 | 0.2670 | 0.04 | 0.2973 | 0.03 |
| **shallow** | FDN | 1.0 | 0.4393 | 0.04 | 0.1817 | 0.03 | 0.2484 | 0.03 |
| **shallow** | FS | 0.1 | 0.1451 | 0.07 | 0.8647 | 0.04 | 0.2455 | 0.05 |
| **shallow** | FS | 0.5 | 0.2080 | 0.02 | 0.4693 | 0.04 | 0.2814 | 0.03 |
| **shallow** | FS | 1.0 | 0.4437 | 0.05 | 0.1731 | 0.06 | 0.2432 | 0.05 |
| **shallow** | FSDN | 0.1 | 0.1496 | 0.06 | 0.8673 | 0.04 | 0.2527 | 0.04 |
| **shallow** | FSDN | 0.5 | 0.4537 | 0.02 | 0.5782 | 0.04 | 0.4993 | 0.03 |
| **shallow** | FSDN | 1.0 | 0.4393 | 0.05 | 0.1817 | 0.04 | 0.2484 | 0.03 |
| **shallow** | SDN | 0.1 | 0.1057 | 0.08 | 0.8796 | 0.04 | 0.1866 | 0.04 |
| **shallow** | SDN | 0.5 | 0.1686 | 0.01 | 0.6982 | 0.05 | 0.2687 | 0.01 |
| **shallow** | SDN | 1.0 | 0.5177 | 0.01 | 0.3267 | 0.06 | 0.3957 | 0.02 |

# Effectiveness: fscore



# Training set size




There are no significant differences in performances with training set ranging from 30 to 100

fscore is a solid optimization measure

# Conclusions

# Conclusions

- Good performances on average on the test data

- Small training set required = small effort for the data creators/curators

- Handle EAD files heterogeneity within the same collection

# Open questions

- Is the achieved effectiveness *enough* for the archivists?

- Is the system solid if tested across collections (*transfer learning*)?

- Is it possible to extend the system to build citations for multiple elements?

# Future directions

- Data citation indexes: we need a method to recognize groups of citations and relate them to the same dataset

- Define and determine citation identity

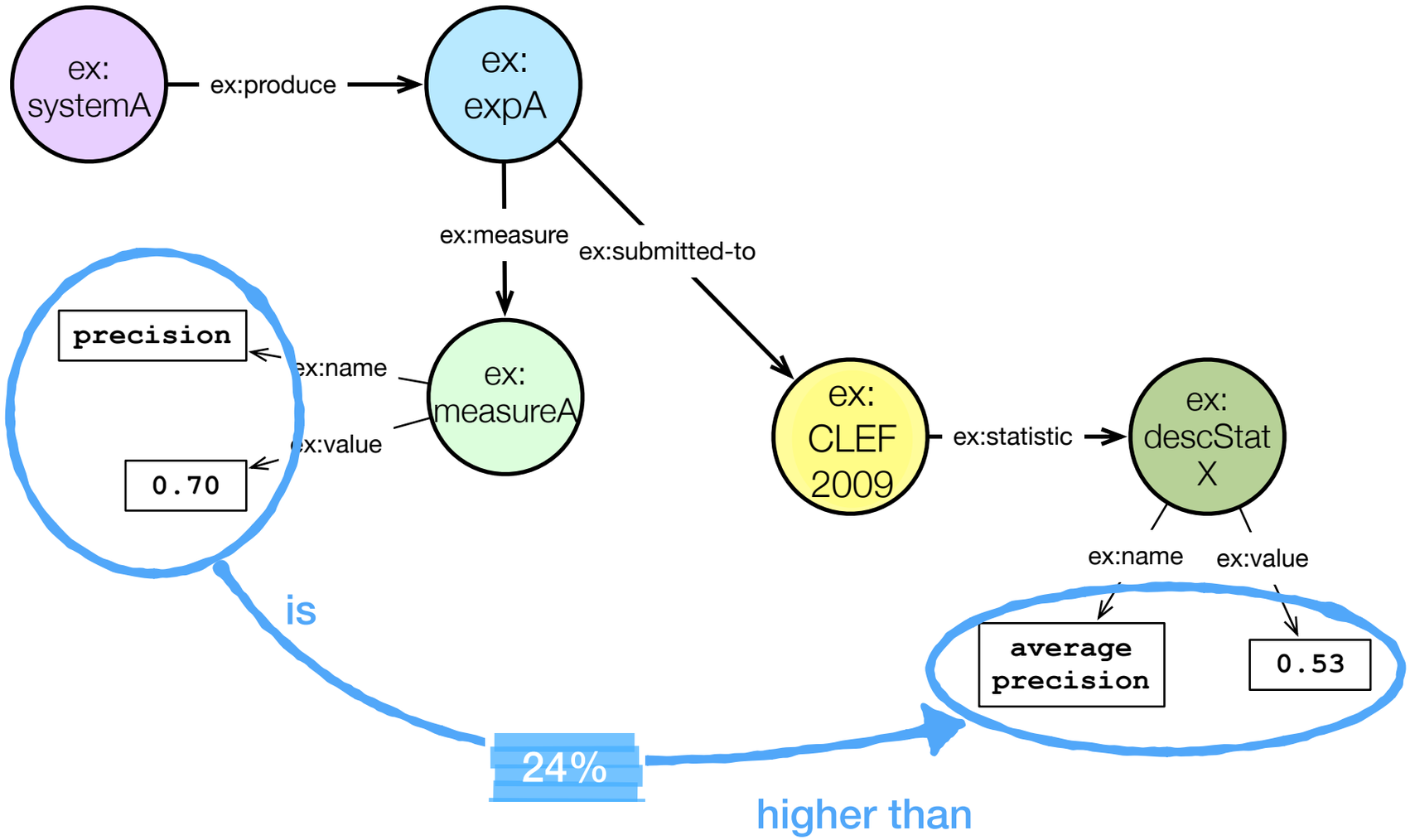
- Beyond XML: What happens with relational databases which cannot be represented as a hierarchy?

- Beyond XML: What happens with graphs (e.g., RDF)?

- Supporting claims…

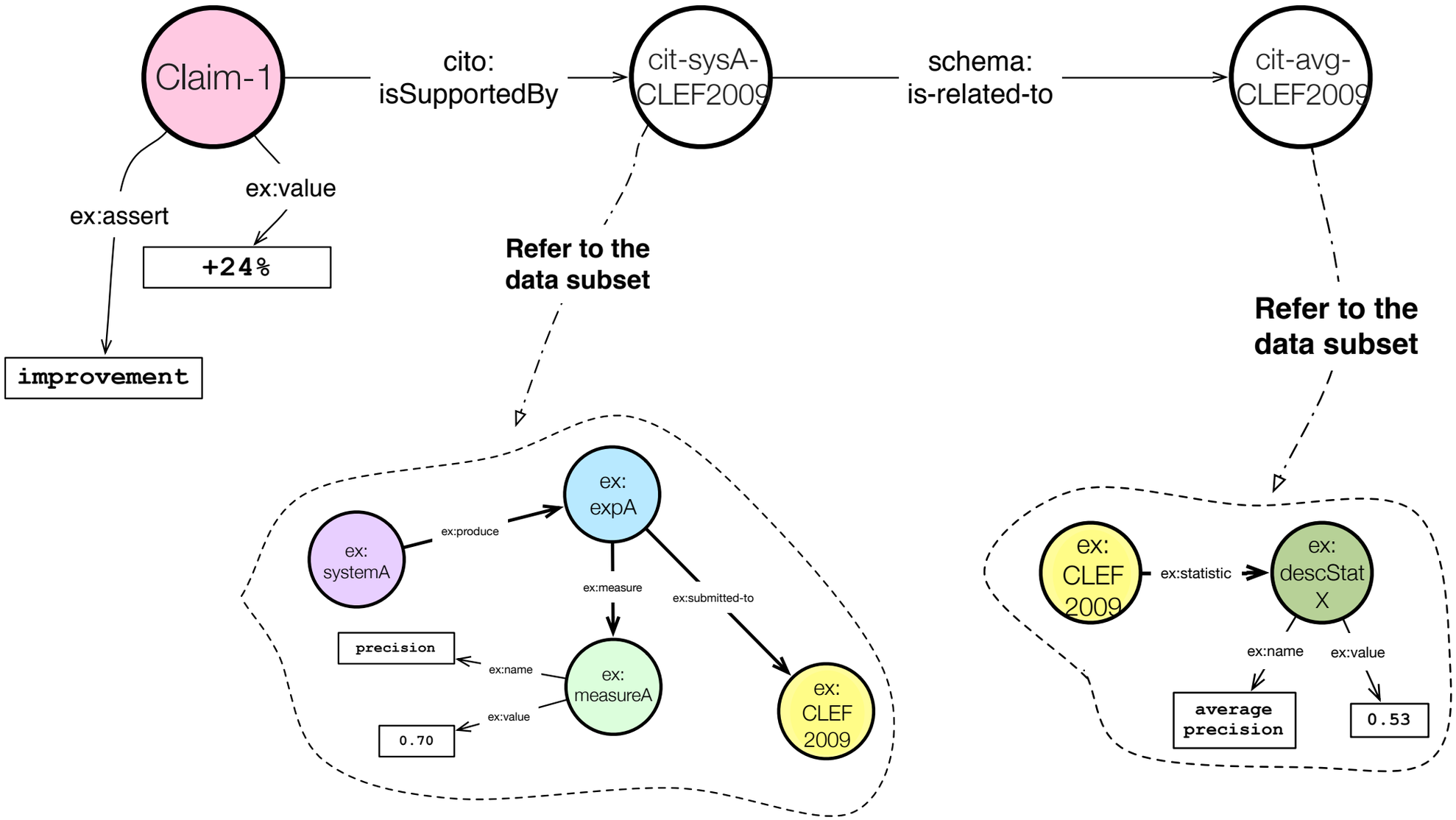
# Future directions: Supporting a claim

"Precision of system A is 24% higher than the average precision of systems which participated in CLEF 2009"



# Future directions: Supporting a claim

"Precision of system A is 24% higher than the average precision of systems which participated in CLEF 2009"



# Questions?