“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
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?

- 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).

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

- 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 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

- 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
From the computer scientists perspective

- 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?

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**.
Encoding of archival data: EAD

(a) Archival Tree

(b) EAD representation
Characteristics of EAD files

- A single 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

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

AH 2005: UK Archival Hub, 2005 snapshot
IISG 2005: International Institute of Social History, 2005 snapshot
NA 2008: Nationaal Archief, The Netherlands, 2008 snapshot
UniMa 2014: University of Maryland, 2014 snapshot
Correspondence, 1951-1956,


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

<table>
<thead>
<tr>
<th>Human-Readable Citation</th>
<th>Machine-Readable Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://hdl.loc.gov/loc.mss/eadmss.ms001024">http://hdl.loc.gov/loc.mss/eadmss.ms001024</a></td>
<td>/ead/eadheader/eadid</td>
</tr>
<tr>
<td>Manuscript Division, Library of Congress</td>
<td>/ead/eadheader/filedesc/publicationstmt/publisher</td>
</tr>
<tr>
<td>Huntington Cairns Papers</td>
<td>/ead/archdesc/did/unittitle</td>
</tr>
<tr>
<td>Part II: Writings</td>
<td>/ead/archdesc/dsc/c01[10]/did/unittitle</td>
</tr>
<tr>
<td>1905-1984</td>
<td>/ead/archdesc/dsc/c01[10]/did/unittitle/unitdate</td>
</tr>
<tr>
<td>box</td>
<td>/ead/archdesc/dsc/c01[10]/did/container/@type</td>
</tr>
<tr>
<td>129-152</td>
<td>/ead/archdesc/dsc/c01[10]/did/container</td>
</tr>
<tr>
<td>By Cairns</td>
<td>/ead/archdesc/dsc/c01[10]/c02[1]/did/unittitle</td>
</tr>
<tr>
<td>box</td>
<td>/ead/archdesc/dsc/c01[10]/c02[1]/did/container/@type</td>
</tr>
<tr>
<td>129</td>
<td>/ead/archdesc/dsc/c01[10]/c02[1]/did/container/</td>
</tr>
<tr>
<td>Books</td>
<td>/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/unittitle</td>
</tr>
<tr>
<td>box</td>
<td>/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container/@type</td>
</tr>
<tr>
<td>135</td>
<td>/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container</td>
</tr>
<tr>
<td>&quot;The Elements of Legal Theory&quot; (unpublished)</td>
<td>/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/did/unittitle</td>
</tr>
<tr>
<td>Correspondence, 1951-1956</td>
<td>/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/c05[1]/did/unittitle</td>
</tr>
</tbody>
</table>
http://hdl.loc.gov/loc.mss/eadmss.ms001024
What does the user see?

Generation of citations: The problem

**Given:**

- **Data:**
  
  /ead/archdesc/dsc/c01[10]/c02/c03[4]/c04[2]/c05[1]/did/unittitle

- **Query (XPath):**

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

- **Dataset:** EAD file

**Generate: 
**

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

1. Human-Readable Citations
   XML Files Collection
   Training Data

2. Learner

3. Citation Model

4. Citation System

5. Citation XPath
   XML File
   Test Data

6. Machine-Readable Citation
   Human-Readable Citation
   Output Reference
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

\[ T = \{t_1, t_2, \ldots, t_n\} \]
\[ \mathcal{H} = \{H_1, H_2, \ldots, H_m\} \]

Matching mode

Citation Parser

\[ \mathcal{H} \]

\[ H_1 = \{h_1, h_2, \ldots\} \]
\[ \ldots \]
\[ H_m = \{h_x, h_y, \ldots\} \]

Textual tokens

XML Retrieval System

Matching paths

Citation Model

Citation Model Builder

Learner
Learner: citation 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

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

- 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.
sample citation model created from a single human-readable citation
Learning to cite framework

1. Human-Readable Citations
2. XML Files Collection
   Training Data
3. Learner
4. Citation System
5. Citation XPath
   XML File
   Test Data
6. Machine-Readable Citation
   Human-Readable Citation
   Output Reference
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?
- 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
System: Pathfinder (example)

test path
/ead/archdesc/dsc/c01

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/c01/did/unittitle

exact match + descendants
- 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**
- 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 to 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`

$relDepth = 2$

We need to rank the candidate paths and keep only the most relevant ones.
System: Path filter (ranking function)

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

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

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

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

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)

<table>
<thead>
<tr>
<th>Candidate paths</th>
<th>FDSN score</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ead/archdesc/dsc/c01/did/unittitle</td>
<td>0.459</td>
</tr>
<tr>
<td>/ead/archdesc/dsc/c01/did/unittitle/unitdate</td>
<td>0.625</td>
</tr>
<tr>
<td>/ead/archdesc/dsc/c01/did/container</td>
<td>0.468</td>
</tr>
<tr>
<td>/ead/archdesc/dsc/c01/did/container/@type</td>
<td>1.000</td>
</tr>
<tr>
<td>/ead/archdesc/dsc/c01/c02/did/container</td>
<td>0.333</td>
</tr>
<tr>
<td>/ead/archdesc/dsc/c01/c02/did/container/@type</td>
<td>0.750</td>
</tr>
<tr>
<td>/ead/archdesc/dsc/c01/c02/did/unittitle</td>
<td>0.310</td>
</tr>
</tbody>
</table>
## System: Path filter (example)

<table>
<thead>
<tr>
<th>Candidate paths</th>
<th>FDSN score</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ead/archdesc/dsc/c01/did/container/@type</td>
<td>1.000</td>
</tr>
<tr>
<td>/ead/archdesc/dsc/c01/c02/did/container/@type</td>
<td>0.750</td>
</tr>
<tr>
<td>/ead/archdesc/dsc/c01/did/unittitle/unitdate</td>
<td>0.625</td>
</tr>
<tr>
<td>/ead/archdesc/dsc/c01/did/container</td>
<td>0.468</td>
</tr>
<tr>
<td>/ead/archdesc/dsc/c01/did/unittitle</td>
<td>0.459</td>
</tr>
<tr>
<td>/ead/archdesc/dsc/c01/c02/did/container/@type</td>
<td>0.333</td>
</tr>
<tr>
<td>/ead/archdesc/dsc/c01/c02/did/unittitle</td>
<td>0.310</td>
</tr>
</tbody>
</table>

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*.

\[
\text{precision} = \frac{\text{correct paths returned}}{\text{total paths returned}}
\]

\[
\text{recall} = \frac{\text{correct paths returned}}{\text{total correct paths}}
\]

\[
\text{fscore} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]
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

<table>
<thead>
<tr>
<th>Tree Type</th>
<th>Ranking Function</th>
<th>Score Threshold</th>
<th>Avg Precision</th>
<th>Std Precision</th>
<th>Avg Recall</th>
<th>Std Recall</th>
<th>Avg Fscore</th>
<th>Std Fscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>exact</td>
<td>FDN</td>
<td>0.1</td>
<td>0.3789</td>
<td>0.06</td>
<td>0.8975</td>
<td>0.04</td>
<td>0.5231</td>
<td>0.04</td>
</tr>
<tr>
<td>exact</td>
<td>FDN</td>
<td>0.5</td>
<td>0.7356</td>
<td>0.01</td>
<td>0.7448</td>
<td>0.03</td>
<td><strong>0.7316</strong></td>
<td>0.01</td>
</tr>
<tr>
<td>exact</td>
<td>FDN</td>
<td>1.0</td>
<td><strong>0.7908</strong></td>
<td>0.04</td>
<td>0.4552</td>
<td>0.05</td>
<td>0.5072</td>
<td>0.04</td>
</tr>
<tr>
<td>exact</td>
<td>FS</td>
<td>0.1</td>
<td>0.3813</td>
<td>0.07</td>
<td>0.8962</td>
<td>0.03</td>
<td>0.5196</td>
<td>0.04</td>
</tr>
<tr>
<td>exact</td>
<td>FS</td>
<td>0.5</td>
<td>0.6042</td>
<td>0.01</td>
<td>0.6919</td>
<td>0.03</td>
<td>0.6372</td>
<td>0.01</td>
</tr>
<tr>
<td>exact</td>
<td>FS</td>
<td>1.0</td>
<td>0.7211</td>
<td>0.02</td>
<td>0.2949</td>
<td>0.05</td>
<td>0.4087</td>
<td>0.03</td>
</tr>
<tr>
<td>exact</td>
<td>FSDN</td>
<td>0.1</td>
<td>0.3769</td>
<td>0.06</td>
<td>0.8975</td>
<td>0.04</td>
<td>0.5208</td>
<td>0.04</td>
</tr>
<tr>
<td>exact</td>
<td>FSDN</td>
<td>0.5</td>
<td>0.7293</td>
<td>0.01</td>
<td>0.7440</td>
<td>0.03</td>
<td>0.7278</td>
<td>0.01</td>
</tr>
<tr>
<td>exact</td>
<td>FSDN</td>
<td>1.0</td>
<td><strong>0.7908</strong></td>
<td>0.04</td>
<td>0.4542</td>
<td>0.08</td>
<td>0.5694</td>
<td>0.05</td>
</tr>
<tr>
<td>exact</td>
<td>SDN</td>
<td>0.1</td>
<td>0.1845</td>
<td>0.04</td>
<td>0.9052</td>
<td>0.04</td>
<td>0.3014</td>
<td>0.04</td>
</tr>
<tr>
<td>exact</td>
<td>SDN</td>
<td>0.5</td>
<td>0.2607</td>
<td>0.00</td>
<td>0.7684</td>
<td>0.04</td>
<td>0.3857</td>
<td>0.01</td>
</tr>
<tr>
<td>exact</td>
<td>SDN</td>
<td>1.0</td>
<td>0.3564</td>
<td>0.01</td>
<td>0.3411</td>
<td>0.04</td>
<td>0.3411</td>
<td>0.02</td>
</tr>
<tr>
<td>mixed</td>
<td>FDN</td>
<td>0.1</td>
<td>0.3186</td>
<td>0.05</td>
<td>0.8942</td>
<td>0.04</td>
<td>0.4631</td>
<td>0.04</td>
</tr>
<tr>
<td>mixed</td>
<td>FDN</td>
<td>0.5</td>
<td>0.5957</td>
<td>0.02</td>
<td>0.7111</td>
<td>0.05</td>
<td>0.6403</td>
<td>0.03</td>
</tr>
<tr>
<td>mixed</td>
<td>FDN</td>
<td>1.0</td>
<td>0.6115</td>
<td>0.04</td>
<td>0.3636</td>
<td>0.04</td>
<td>0.4477</td>
<td>0.03</td>
</tr>
<tr>
<td>mixed</td>
<td>FS</td>
<td>0.1</td>
<td>0.3339</td>
<td>0.08</td>
<td>0.8901</td>
<td>0.05</td>
<td>0.4734</td>
<td>0.06</td>
</tr>
<tr>
<td>mixed</td>
<td>FS</td>
<td>0.5</td>
<td>0.6127</td>
<td>0.03</td>
<td>0.6473</td>
<td>0.04</td>
<td>0.6220</td>
<td>0.03</td>
</tr>
<tr>
<td>mixed</td>
<td>FS</td>
<td>1.0</td>
<td>0.7028</td>
<td>0.04</td>
<td>0.2990</td>
<td>0.10</td>
<td>0.4095</td>
<td>0.06</td>
</tr>
<tr>
<td>mixed</td>
<td>FSDN</td>
<td>0.1</td>
<td>0.3276</td>
<td>0.05</td>
<td>0.8942</td>
<td>0.04</td>
<td>0.4718</td>
<td>0.04</td>
</tr>
<tr>
<td>mixed</td>
<td>FSDN</td>
<td>0.5</td>
<td>0.6514</td>
<td>0.02</td>
<td>0.7252</td>
<td>0.05</td>
<td>0.6789</td>
<td>0.03</td>
</tr>
<tr>
<td>mixed</td>
<td>FSDN</td>
<td>1.0</td>
<td>0.7746</td>
<td>0.03</td>
<td>0.4472</td>
<td>0.05</td>
<td>0.5584</td>
<td>0.04</td>
</tr>
<tr>
<td>mixed</td>
<td>SDN</td>
<td>0.1</td>
<td>0.1469</td>
<td>0.05</td>
<td><strong>0.9045</strong></td>
<td>0.04</td>
<td>0.2493</td>
<td>0.05</td>
</tr>
<tr>
<td>mixed</td>
<td>SDN</td>
<td>0.5</td>
<td>0.2690</td>
<td>0.01</td>
<td>0.7676</td>
<td>0.05</td>
<td>0.3948</td>
<td>0.01</td>
</tr>
<tr>
<td>mixed</td>
<td>SDN</td>
<td>1.0</td>
<td>0.4234</td>
<td>0.01</td>
<td>0.3643</td>
<td>0.05</td>
<td>0.3822</td>
<td>0.02</td>
</tr>
<tr>
<td>shallow</td>
<td>FDN</td>
<td>0.1</td>
<td>0.1630</td>
<td>0.04</td>
<td>0.8679</td>
<td>0.04</td>
<td>0.2719</td>
<td>0.04</td>
</tr>
<tr>
<td>shallow</td>
<td>FDN</td>
<td>0.5</td>
<td>0.3645</td>
<td>0.02</td>
<td>0.2670</td>
<td>0.04</td>
<td>0.2973</td>
<td>0.03</td>
</tr>
<tr>
<td>shallow</td>
<td>FDN</td>
<td>1.0</td>
<td>0.4393</td>
<td>0.04</td>
<td>0.1817</td>
<td>0.03</td>
<td>0.2484</td>
<td>0.03</td>
</tr>
<tr>
<td>shallow</td>
<td>FS</td>
<td>0.1</td>
<td>0.1451</td>
<td>0.07</td>
<td>0.8647</td>
<td>0.04</td>
<td>0.2455</td>
<td>0.05</td>
</tr>
<tr>
<td>shallow</td>
<td>FS</td>
<td>0.5</td>
<td>0.2080</td>
<td>0.02</td>
<td>0.4693</td>
<td>0.04</td>
<td>0.2814</td>
<td>0.03</td>
</tr>
<tr>
<td>shallow</td>
<td>FS</td>
<td>1.0</td>
<td>0.4437</td>
<td>0.05</td>
<td>0.1731</td>
<td>0.06</td>
<td>0.2432</td>
<td>0.05</td>
</tr>
<tr>
<td>shallow</td>
<td>FSDN</td>
<td>0.1</td>
<td>0.1496</td>
<td>0.06</td>
<td>0.8673</td>
<td>0.04</td>
<td>0.2527</td>
<td>0.04</td>
</tr>
<tr>
<td>shallow</td>
<td>FSDN</td>
<td>0.5</td>
<td>0.4537</td>
<td>0.02</td>
<td>0.5782</td>
<td>0.04</td>
<td>0.4993</td>
<td>0.03</td>
</tr>
<tr>
<td>shallow</td>
<td>FSDN</td>
<td>1.0</td>
<td>0.4393</td>
<td>0.05</td>
<td>0.1817</td>
<td>0.04</td>
<td>0.2484</td>
<td>0.03</td>
</tr>
<tr>
<td>shallow</td>
<td>SDN</td>
<td>0.1</td>
<td>0.1057</td>
<td>0.08</td>
<td>0.8796</td>
<td>0.04</td>
<td>0.1866</td>
<td>0.04</td>
</tr>
<tr>
<td>shallow</td>
<td>SDN</td>
<td>0.5</td>
<td>0.1686</td>
<td>0.01</td>
<td>0.6982</td>
<td>0.05</td>
<td>0.2687</td>
<td>0.01</td>
</tr>
<tr>
<td>shallow</td>
<td>SDN</td>
<td>1.0</td>
<td>0.5177</td>
<td>0.01</td>
<td>0.3267</td>
<td>0.06</td>
<td>0.3957</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Effectiveness: fscore

Fscore averaged over 10 training set random samples as the size varies

Fscore

Training Set Sample 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?