Label Annotation through Biodiversity Enhanced Learning

P. Bryan Heidorn
University of Arizona
pbryan.heidorn@gmail.com

Qianjin Zhang
University of Arizona
zhqjnj@gmail.com

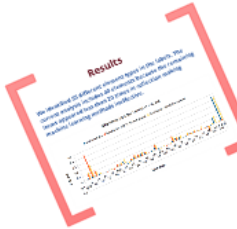
Acknowledgment

- Steven Chong's contribution to enhance training data
- The University of Florida, Biological Science Collections Tracker
- School of Information Resource & Library Science, the University of Arizona
- The University of Arizona
- Hong Cui and Sunali Ranade's suggestion

Results

Hay! it does not work 100% of the time.

Results



My technique works better on my tiny museum labels (collected between 2009-2011 on my keyboard) on my Commodore 64.

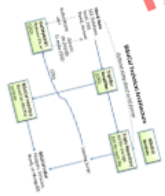
5-fold evaluation for training sets

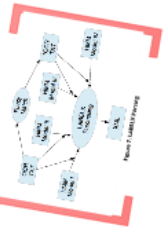
- The control
  - 150 original labels from Tyler and 200 original labels from Museum of the North (350 original labels)
- Enhancement
  - 350 original labels + 350 corrected labels without OCR errors
- Substitution with and Enhancement
  - 350 original labels + 350 labels with substitution
- Substitution with Enhancement
  - 350 original labels + 350 corrected labels without OCR errors + 350 labels with substitution

Why museums, scientific names, foreigners, and barcodes?

See how the scientific names, collector and locality authority files

XML To RDF





LABELX Workflow

Hay! You spelled Dublin Core wrong

The Post Office did this already.

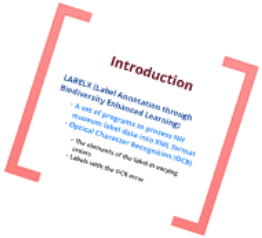
Ideal XML markup of the OCRed label

Background

Introduction

LABELX (Label Annotation through Biodiversity Enhanced Learning)

- A set of programs to process bar museum label data into XML format
- Optical Character Recognition (OCR)



Label Annotation through Biodiversity Enhanced Learning

P. Bryan Heidorn
University of Arizona
pbryan.heidorn@gmail.com

Qianjin Zhang
University of Arizona
zhqjnj@gmail.com

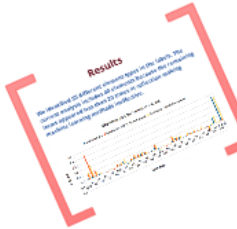
Acknowledgment

- Steven Chong's contribution to enhance training data
- The University of Florida, Biological Science Collections Tracker
- School of Information Resource & Library Science, the University of Arizona
- The University of Arizona
- Hong Cui and Sunali Ranade's suggestion

Results

Hay! it does not work 100% of the time.

Results



My technique works better on my tiny museum labels (collected between 2009-2011 on my keyboard) on my Commodore 64.

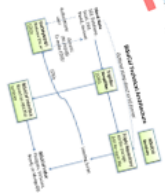
5-fold evaluation for training sets

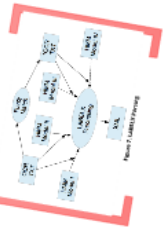
- The control
  - 150 original labels from Tyler and 200 original labels from Museum of the North (350 original labels)
- Enhancement
  - 350 original labels + 350 corrected labels without OCR errors
- Substitution with and Enhancement
  - 350 original labels + 350 labels with substitution
- Substitution with Enhancement
  - 350 original labels + 350 corrected labels without OCR errors + 350 labels with substitution

Why museums, scientific names, foreigners, and barcodes?

See how the scientific names, collector and locality authority files

XML To RDF





LABELX Workflow

Hay! You spelled Dublin Core wrong

The Post Office did this already.

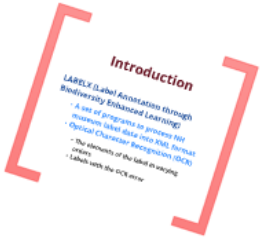
Ideal XML markup of the OCRed label

Background

Introduction

LABELX (Label Annotation through Biodiversity Enhanced Learning)

- A set of programs to process bar museum label data into XML format
- Optical Character Recognition (OCR)



Introduction

LABELX (Label Annotation through Biodiversity Enhanced Learning)

- A set of programs to process NH museum label data into XML format
- Optical Character Recognition (OCR)
  - The elements of the label in varying orders
  - Labels with the OCR error

Background

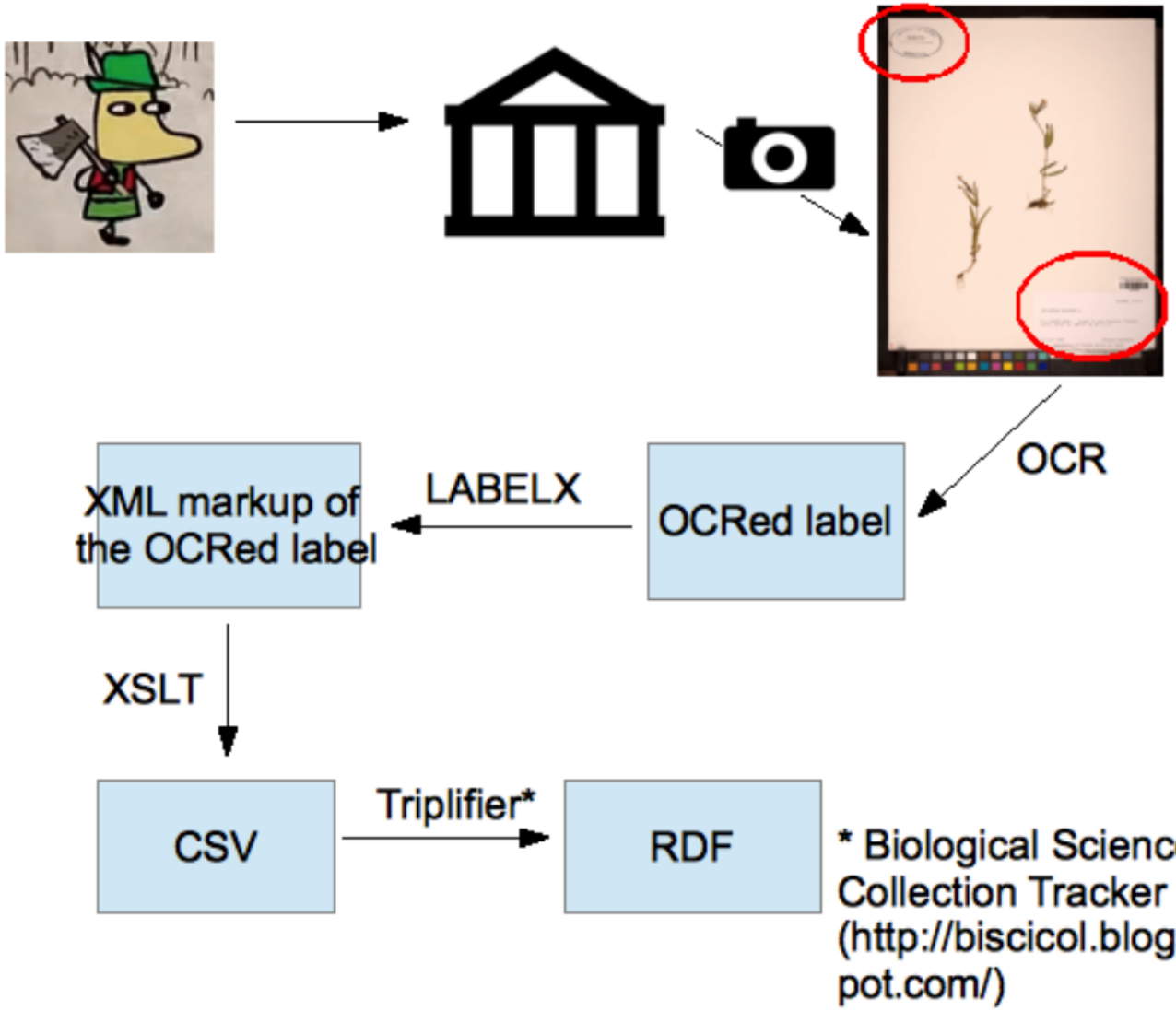

OCR
OCRed label
LABELX
XML markup of the OCRed label
XSLT
CSV
Triplifier*
RDF
* Biological Science Collection Tracker (http://biscicol.blogspot.com/)

Figure 2.
Example Museum Specimen Label

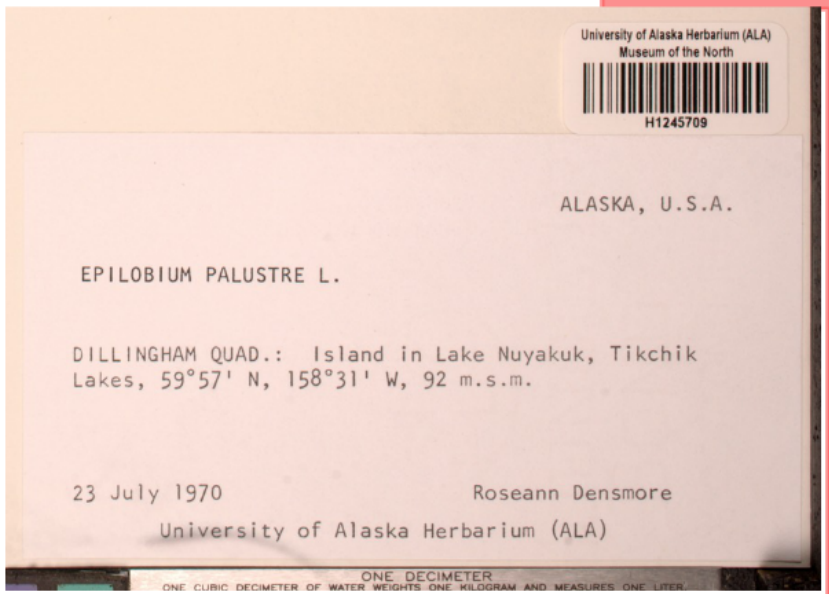
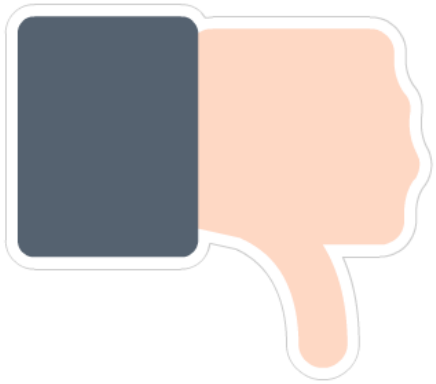


Figure 3. Example Label

University of Alaska Herbarium (ALA)
Museum of the North
H1245709
ALASKA, U.S.A.
EPILOBIUM PALUSTRE L.
DILLINGHAM QUAD.: Island in Lake Nuyakuk, Tikchik Lakes, 59°§7' N, l58°3i' W, 92 m.s.m.
23 July 1970 Roseann Densmore
University of Alaska Herbarium (ALA)

Figure 4. OCRed label

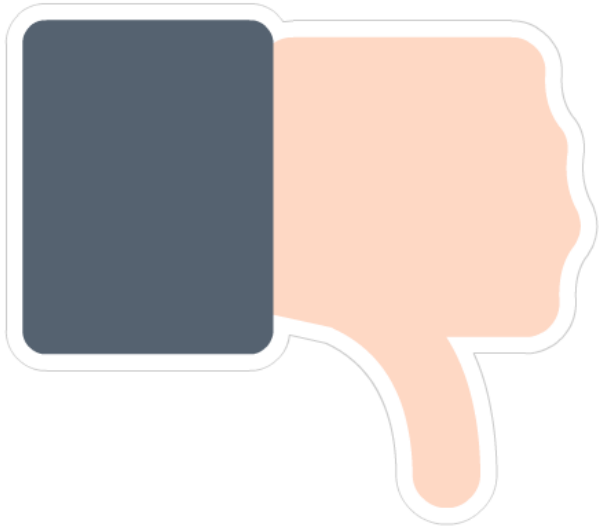


The Post Office did this already.

Ideal XML markup of the OCRed label

```
<labeldata>
<globally unique identifier>Uniscicol0001254A</globally unique identifier>
<institution>University of Alaska Herbarium (ALA)</institution>
<institution>Museum of the North</institution>
<barcode>H1245709</barcode>
<institutionLocation>ALASKA, U.S.A.</institutionLocation>
<genus>EPILOBIUM</genus>
<specificEpithet>PALUSTRE</specificEpithet>
<scientificNameAuthorship>L.</scientificNameAuthorship>
<locality>DILLINGHAM QUAD.: Island in Lake Nuyakuk, Tikchik
Lakes,</locality>
<verbatimCoordinates>59°§7&apos; N, I58°3i&apos; W, 92
m.s.m.</verbatimCoordinates>
<collectionDate>23 July 1970</collectionDate>
<collector>Roseann Densmore</collector>
<institution>University of Alaska Herbarium (ALA)</institution>
</labeldata>
```

Figure 5. Example of XML markup of the OCRed label



Hay! You spelled Dublin Core wrong

LABELX Workflow

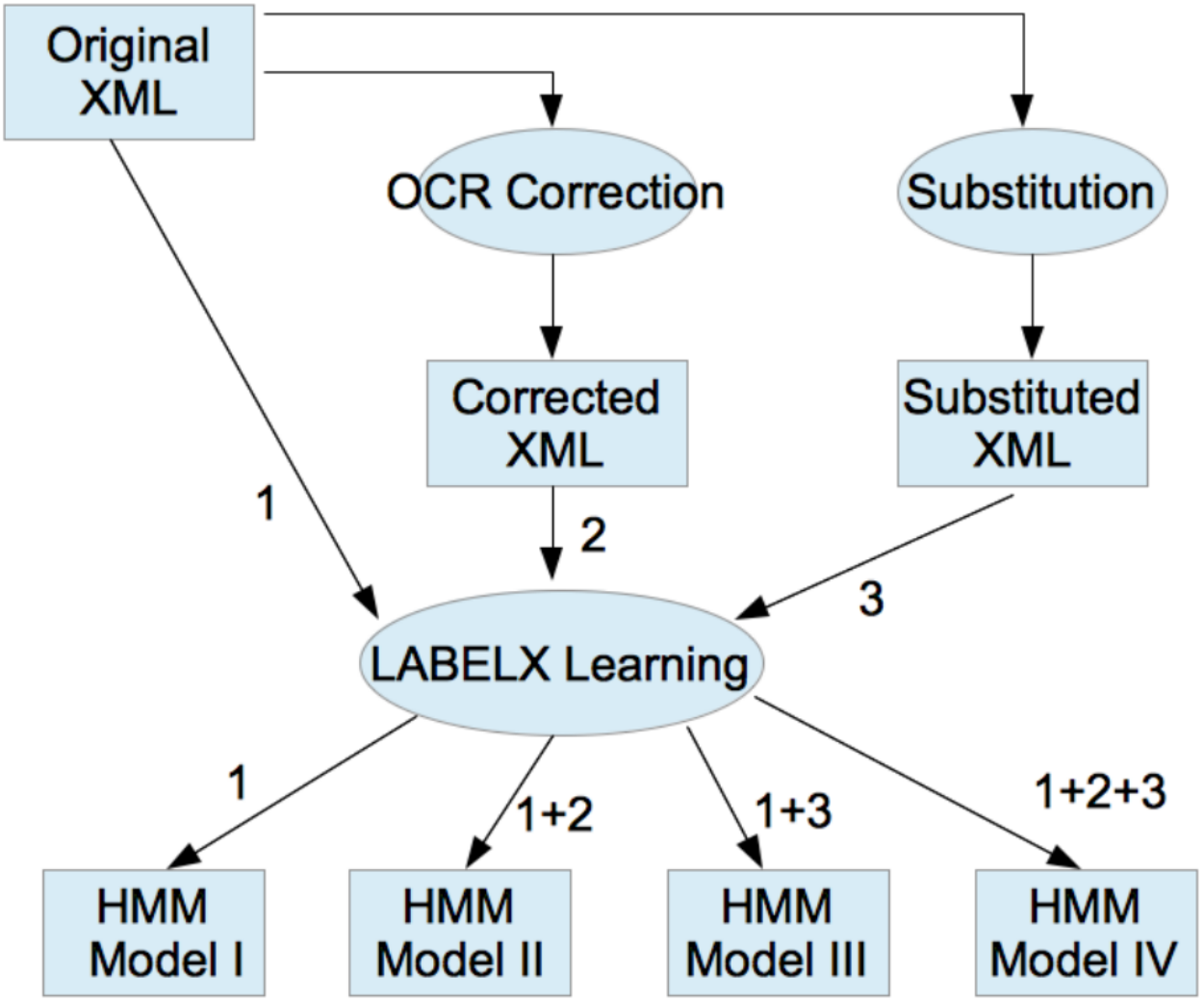


Figure 6. LABELX Learning

Critical Random Graphs, Regular expressions, simulated annealing or a Harry Potter spell would work better.

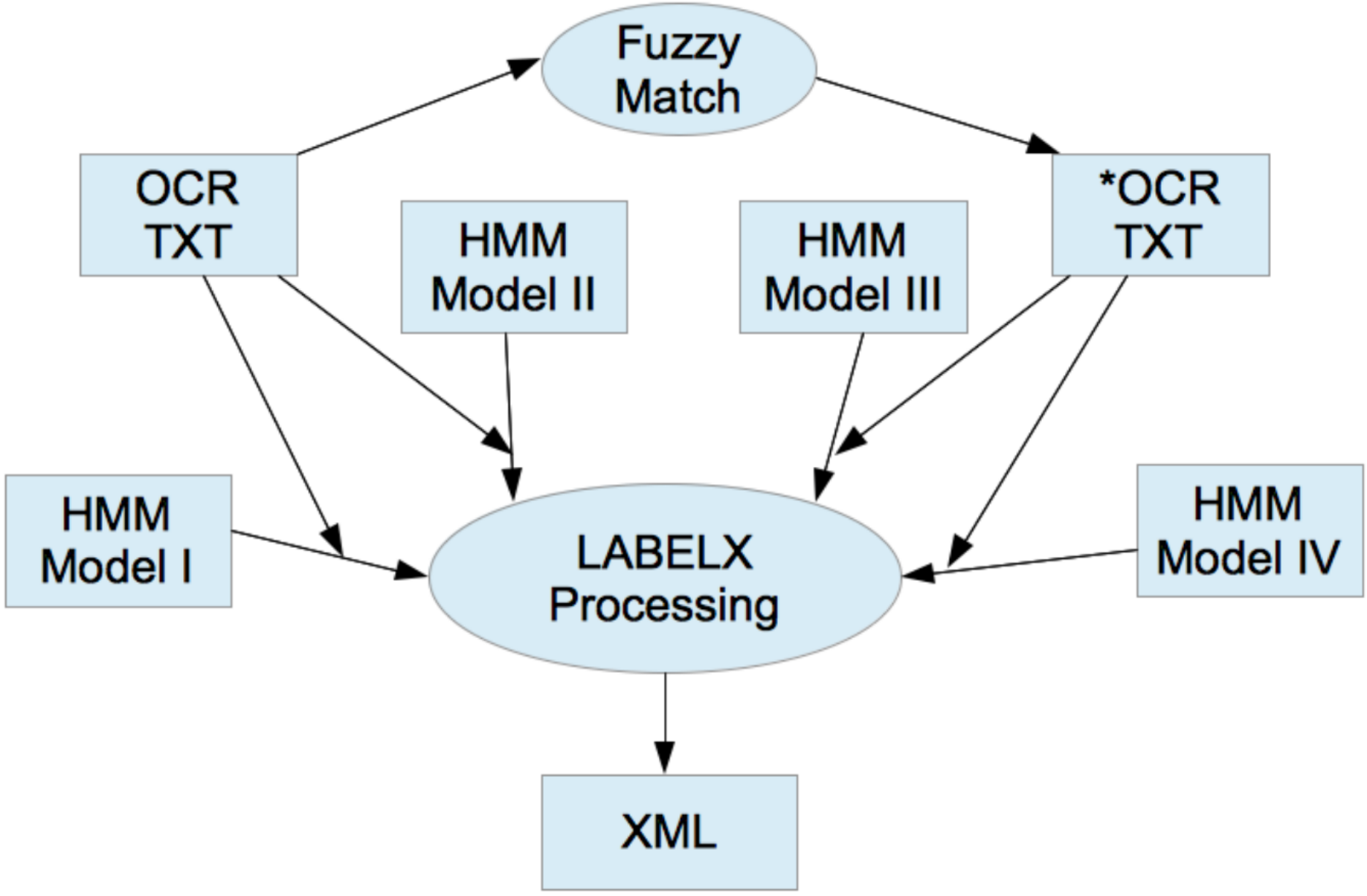
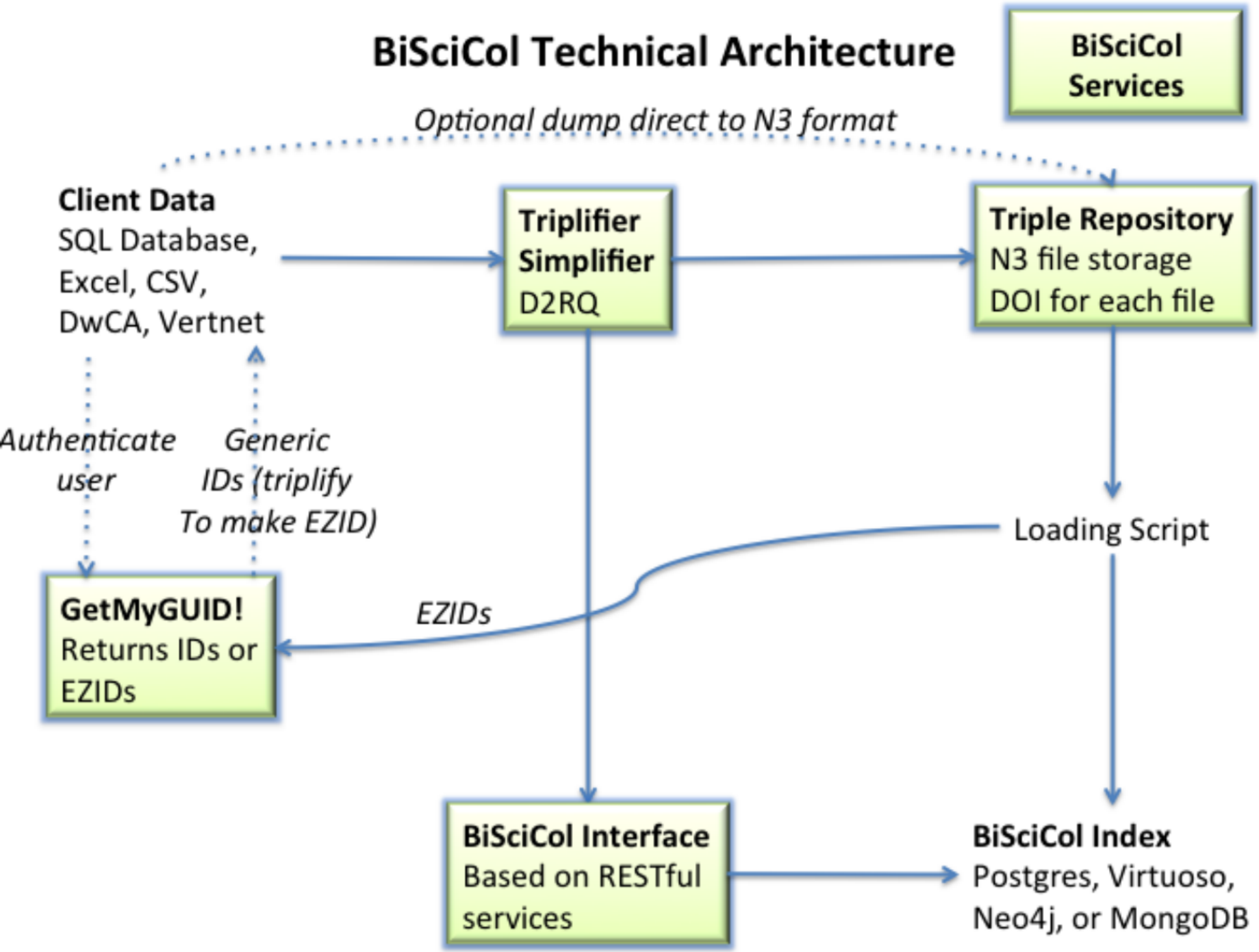


Figure 7. LABELX Parsing

BiSciCol Relationship Terms

RDF	Definition	Example	Symmetric	Transitive
obj1 —ro:derives_from→ obj2	Physical material (obj1) that is substantially derived from other physical material (obj2)	Tissue (obj1) derived from a Specimen (obj2)		Yes
obj1 ⋯biscicol:depends_on⋯▶ obj2	An entity (obj1) whose existence depends on another entity (obj2)	Specimen (obj1) part of a collecting event instance (obj2)		
obj1 ⟷biscicol:alias_of⟷ obj2	Two instances (obj1,obj2) that are understood to be the same thing	Obj1 and Obj2 both refer to the same specimen.	Yes	Yes
obj1 ◀⋯biscicol:related_to⋯▶ obj2	An entity (obj1) whose existence does not solely depend on another entity (obj2)	Agent/Person (obj1) related to Agent/ Institution (obj2).	Yes	

BiSciCol Technical Architecture



XML To RDF

<urn:x-biscicol:RecordUniversity of Alaska Herbarium (ALA)> <http://www.w3.org/ns/ma-ont#isRelatedTo> <urn:x-biscicol:RecordALASKA, U.S.A.> .
<urn:x-biscicol:RecordALASKA, U.S.A.> <http://rs.tdwg.org/dwc/terms/locationAttributes> "University of Alaska Herbarium (ALA)" .
<urn:x-biscicol:RecordALASKA, U.S.A.> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://rs.tdwg.org/dwc/terms/Dataset> .
<urn:x-biscicol:RecordUniversity of Alaska Herbarium (ALA)> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://rs.tdwg.org/dwc/terms/Dataset> .
<urn:x-biscicol:TaxonEPILOBIUM> <http://www.w3.org/ns/ma-ont#isRelatedTo> <urn:x-biscicol:TaxonPALUSTRE> .
<urn:x-biscicol:TaxonPALUSTRE> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://rs.tdwg.org/dwc/terms/Taxon> .
<urn:x-biscicol:TaxonEPILOBIUM> <http://rs.tdwg.org/dwc/terms/higherTaxonConceptID> "PALUSTRE" .
<urn:x-biscicol:TaxonEPILOBIUM> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://rs.tdwg.org/dwc/terms/Taxon> .
<urn:x-biscicol:EventRoseann Densmore> <http://www.w3.org/ns/ma-ont#isSourceOf> <urn:x-biscicol:Event1970-07-23T00:00:00.000-06:00> .
<urn:x-biscicol:EventRoseann Densmore> <http://www.w3.org/ns/ma-ont#isRelatedTo> <urn:x-biscicol:EventDILLINGHAM QUAD.: Island in Lake Nuyakuk, Tikchik Lakes,> .
<urn:x-biscicol:EventDILLINGHAM QUAD.: Island in Lake Nuyakuk, Tikchik Lakes,> <http://rs.tdwg.org/dwc/terms/locationAccordingTo> "Roseann Densmore" .
<urn:x-biscicol:EventDILLINGHAM QUAD.: Island in Lake Nuyakuk, Tikchik Lakes,> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://rs.tdwg.org/dwc/terms/Event> .

Figure 8. Triples (Selected) from Triplifier

Why generalize scientific names, integers, and barcode?

- Scientific names are difficult for HMM because they are rare.
- So are integers and barcodes.

Why do fuzzy-match for scientific names 

- Simple substitution leads to false positives.
  - Genus names and species names are not always unique.
- OCR errors leads to false negatives.

Just use the scientific name, collector and locality authority files!

- **Fuzzy-match for scientific names: replace genus name w/ BiscicolGenus; species name with BiscicolSpecies**

H1173189.txt:::<gn>PHYSARIA:::<sp cc = "acutifolia">ACUTIFDLIA
H1146037.txt:::<gn cc = "luzula">LUVZULA:::<sp>MULTIFLORA

- **Integer identification: replace integers excluding range from 1900 to 2012 and from 1 to 31 w/ BiscicolInt**
- **Specific match for barcodes: replace barcodes w/ AlphaInt**
- **Collection dates: replace w/ BiscicolDate**

University of Alaska Herbarium (ALA)
Museum of the North
H1245709
ALASKA, U.S.A.
EPILOBIUM PALUSTRE L.
DILLINGHAM QUAD.: Island in Lake Nuyakuk, Tikchik
Lakes, 59°§7' N, l58°3i' W, 92 m.s.m.
23 July 1970 Roseann Densmore
University of Alaska Herbarium (ALA)

Before Substitution

University of Alaska Herbarium (ALA)
Museum of the North
AlphaInt
ALASKA, U.S.A.
BiscicolGenus BiscicolSpecies L.
DILLINGHAM QUAD.: Island in Lake Nuyakuk, Tikchik
Lakes, 59°§7' N, l58°3i' W, BiscicolInt m.s.m.
BiscicolDate Roseann Densmore
University of Alaska Herbarium (ALA)

After Substitution

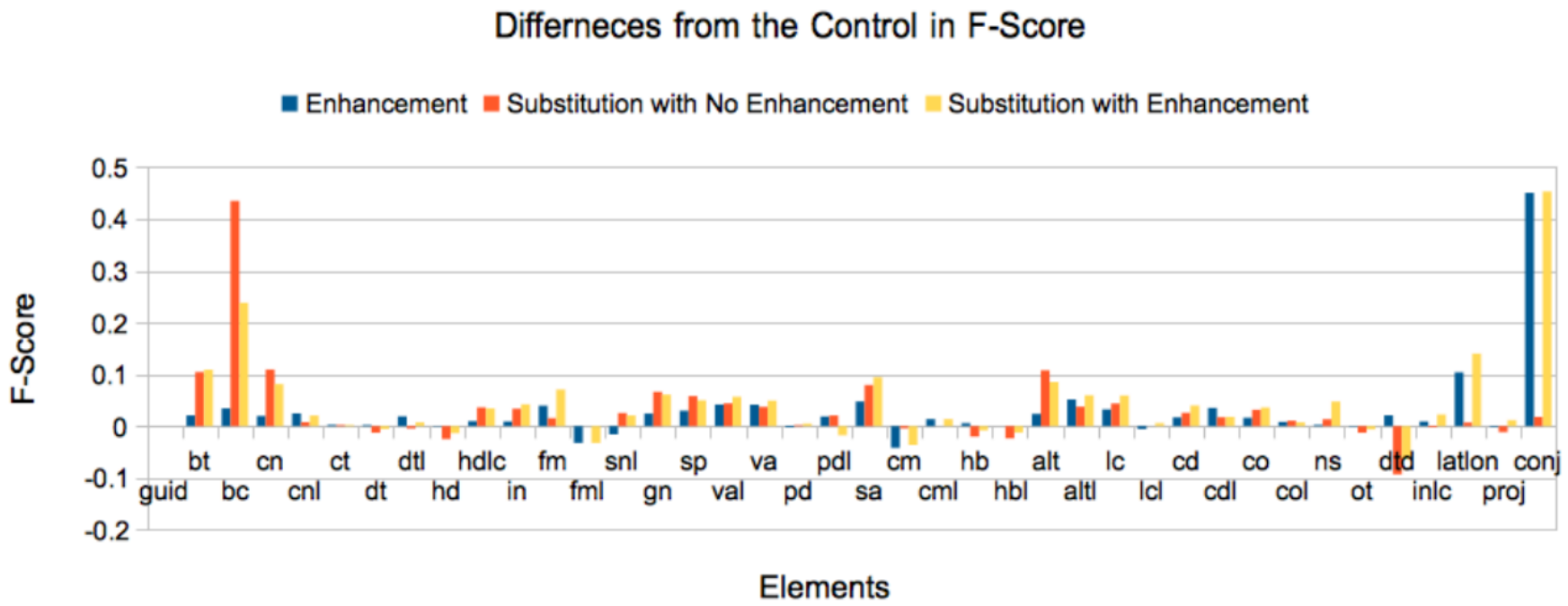
5-fold evaluation for training sets

- The control
  - 130 original labels from Yale and 200 original labels from Museum of the North (330 original labels)
- Enhancement
  - 330 original labels + 330 corrected labels without OCR errors
- Substitution without Enhancement
  - 330 original labels + 330 labels with substitution
- Substitution with Enhancement
  - 330 original labels + 330 corrected labels without OCR errors + 330 labels with substitution

My technique works better on my 100 museum labels (collected between 2009-2011 in my backyard) on my Commodor 64.

Results

We identified 55 different element types in the labels. The current analysis includes 40 elements because the remaining items appeared less than 20 times in collection making machine learning methods ineffective.



Hay! It does not work 100% of the time.

Results

Elements	Control	Sub+Enhanced	Dif (%)	Elements	Control	Sub+Enhanced	Dif (%)
bc	0.4264	0.6647	+23.82	sa	0.6984	0.7928	+9.44
cn	0.5267	0.6078	+8.11	cm	0.7849	0.7487	-3.63
dt	0.7014	0.6957	-0.58	cml	0.8132	0.8268	+1.36
hd	0.5613	0.5478	-1.34	hb	0.7177	0.7096	-0.81
in	0.8459	0.8879	+4.21	alt	0.4175	0.5029	+8.53
fm	0.6877	07584	+7.06	altl	0.5778	0.6374	+5.96
fml	0.8627	0.8302	-3.25	lc	0.7536	0.8128	+5.92
gn	0.7584	0.8193	+6.09	cd	0.8415	0.8814	+3.98
sp	0.7623	0.8120	+4.97	co	0.7773	0.8133	+3.60
val	0.5833	0.6400	+5.67	dtd	0.6597	0.5957	-6.39
va	0.4628	0.5120	+4.92	latlon	0.2403	0.3802	+13.99

Table. Performance Changes in F-Score

Acknowledgment

- Steven Chong's contribution to enhance training data
- The University of Florida, Biological Science Collections Tracker
- School of Information Resource & Library Science, the University of Arizona
- The University of Arizona
- Hong Cui and Sonali Ranade's suggestion

THE UNIVERSITY OF ARIZONA.