

Qianjin Zhang
University of Arizona
zhqjn@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 Sanali Ranade's suggestion

Results

Study	Sample Size	Effect Size	Power	Significance Level	p-value	Confidence Interval	Standard Error	t-statistic	z-statistic
1	100	0.1	0.8	0.05	0.001	0.05	0.02	2.5	2.5
2	200	0.2	0.9	0.01	0.000	0.02	0.01	3.5	3.5
3	300	0.3	0.95	0.001	0.000	0.01	0.005	4.5	4.5
4	400	0.4	0.99	0.000	0.000	0.005	0.002	5.5	5.5
5	500	0.5	1.0	0.000	0.000	0.002	0.001	6.5	6.5
6	600	0.6	1.0	0.000	0.000	0.001	0.0005	7.5	7.5
7	700	0.7	1.0	0.000	0.000	0.0005	0.0002	8.5	8.5
8	800	0.8	1.0	0.000	0.000	0.0002	0.0001	9.5	9.5
9	900	0.9	1.0	0.000	0.000	0.0001	0.00005	10.5	10.5
10	1000	1.0	1.0	0.000	0.000	0.00005	0.00002	11.5	11.5

High Performance Computing & Analytics

My technique works better on my 100
common labels (collected between
2004-2011 in my body and) on my
Commander 64.

Results

After 10 minutes, Group 1 had answered 10 questions. Group 2 had answered 5 questions. After 20 minutes, Group 1 had answered 10 questions. Group 2 had answered 10 questions.

Number of correct answers

Time (minutes)

Group 1: 10 children

Group 2: 10 children

- **True control**
 - 130 original labels from: Yale and 200 original labels from: Muscovy of the North (334 original labels)
- **Enrichment**
 - 334 original labels + 334 corrected labels without OCR errors
- **Substitution with and without enrichment**
 - 334 original labels + 334 labels with substitution
- **Substitution with and without enrichment**
 - 334 original labels + 333 corrected labels without OCR errors + 334 labels with substitution

[illegible]

Why generalize scientific names, integers, and barcodes?

- Scientific names are difficult for data analysis. One can use integers to store integers and barcodes.

XML To RDF

Just use the electronic name collector and easily achieve this!

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

The diagram illustrates a process flow. At the top, a person icon is connected by an arrow to a building icon, which is then connected to a document icon. A feedback arrow points from the document icon back to the person icon. Below this, a flowchart shows a box labeled '1.1.1.1' connected to a box labeled '1.1.1.2', which is then connected to a box labeled '1.1.1.3'. A feedback arrow points from the '1.1.1.3' box back to the '1.1.1.1' box. To the right of the '1.1.1.3' box, there is a list of items: '1.1.1.3.1.1.1', '1.1.1.3.1.1.2', '1.1.1.3.1.1.3', '1.1.1.3.1.1.4', '1.1.1.3.1.1.5', '1.1.1.3.1.1.6', '1.1.1.3.1.1.7', '1.1.1.3.1.1.8', '1.1.1.3.1.1.9', '1.1.1.3.1.1.10', '1.1.1.3.1.1.11', '1.1.1.3.1.1.12', '1.1.1.3.1.1.13', '1.1.1.3.1.1.14', '1.1.1.3.1.1.15', '1.1.1.3.1.1.16', '1.1.1.3.1.1.17', '1.1.1.3.1.1.18', '1.1.1.3.1.1.19', '1.1.1.3.1.1.20', '1.1.1.3.1.1.21', '1.1.1.3.1.1.22', '1.1.1.3.1.1.23', '1.1.1.3.1.1.24', '1.1.1.3.1.1.25', '1.1.1.3.1.1.26', '1.1.1.3.1.1.27', '1.1.1.3.1.1.28', '1.1.1.3.1.1.29', '1.1.1.3.1.1.30', '1.1.1.3.1.1.31', '1.1.1.3.1.1.32', '1.1.1.3.1.1.33', '1.1.1.3.1.1.34', '1.1.1.3.1.1.35', '1.1.1.3.1.1.36', '1.1.1.3.1.1.37', '1.1.1.3.1.1.38', '1.1.1.3.1.1.39', '1.1.1.3.1.1.40', '1.1.1.3.1.1.41', '1.1.1.3.1.1.42', '1.1.1.3.1.1.43', '1.1.1.3.1.1.44', '1.1.1.3.1.1.45', '1.1.1.3.1.1.46', '1.1.1.3.1.1.47', '1.1.1.3.1.1.48', '1.1.1.3.1.1.49', '1.1.1.3.1.1.50', '1.1.1.3.1.1.51', '1.1.1.3.1.1.52', '1.1.1.3.1.1.53', '1.1.1.3.1.1.54', '1.1.1.3.1.1.55', '1.1.1.3.1.1.56', '1.1.1.3.1.1.57', '1.1.1.3.1.1.58', '1.1.1.3.1.1.59', '1.1.1.3.1.1.60', '1.1.1.3.1.1.61', '1.1.1.3.1.1.62', '1.1.1.3.1.1.63', '1.1.1.3.1.1.64', '1.1.1.3.1.1.65', '1.1.1.3.1.1.66', '1.1.1.3.1.1.67', '1.1.1.3.1.1.68', '1.1.1.3.1.1.69', '1.1.1.3.1.1.70', '1.1.1.3.1.1.71', '1.1.1.3.1.1.72', '1.1.1.3.1.1.73', '1.1.1.3.1.1.74', '1.1.1.3.1.1.75', '1.1.1.3.1.1.76', '1.1.1.3.1.1.77', '1.1.1.3.1.1.78', '1.1.1.3.1.1.79', '1.1.1.3.1.1.80', '1.1.1.3.1.1.81', '1.1.1.3.1.1.82', '1.1.1.3.1.1.83', '1.1.1.3.1.1.84', '1.1.1.3.1.1.85', '1.1.1.3.1.1.86', '1.1.1.3.1.1.87', '1.1.1.3.1.1.88', '1.1.1.3.1.1.89', '1.1.1.3.1.1.90', '1.1.1.3.1.1.91', '1.1.1.3.1.1.92', '1.1.1.3.1.1.93', '1.1.1.3.1.1.94', '1.1.1.3.1.1.95', '1.1.1.3.1.1.96', '1.1.1.3.1.1.97', '1.1.1.3.1.1.98', '1.1.1.3.1.1.99', '1.1.1.3.1.1.100'.

idea) XML markup of the OCRed label

Hay! You spelled
Dublin Core wrong

The Post Office did this already.

Qianjin Zhang
University of Arizona
zhqjn@gmail.com



- 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 Sanati Ranade's suggestion

Production

LABELX (Label Annotation through Biodiversity Enhanced Learning)

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

[illegible][illegible]

Results

The identified 10 different incorrect types in the tests, the correct answers included all answers that were given. The results of the tests show that 10 types of incorrect answers were identified and the correct answer was given.

Percentage of students who were able to identify the correct answer

Number of questions

Correct answer

Incorrect answer

5-fold evaluation for training sets

- The control**
 - 130 original labels from Yale and 390 original labels from Microsoft of 100 words (330 original labels)
- Enrichment**
 - 330 original labels + 330 connected labels without OCR errors
- Substitution without enrichment**
 - 330 original labels + 330 labels with substitution
- Substitution with enrichment**
 - 330 original labels + 330 connected labels without OCR errors + 330 labels with substitution

[illegible]

Why population-specific variants?
 Alleles that are common in one population but rare in another
 → may be important for disease susceptibility

Why population-specific variants?
 Alleles that are common in one population but rare in another
 → may be important for disease susceptibility

[illegible]

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

My technician works better on my 190
muscle labels (collected between
2004-2011) at my backyard) on my
Commander 64.

The Post Office did this already.

Hay! You spelled Dublin Core wrong

Don't use the scriptable name, folder and
family member files!

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

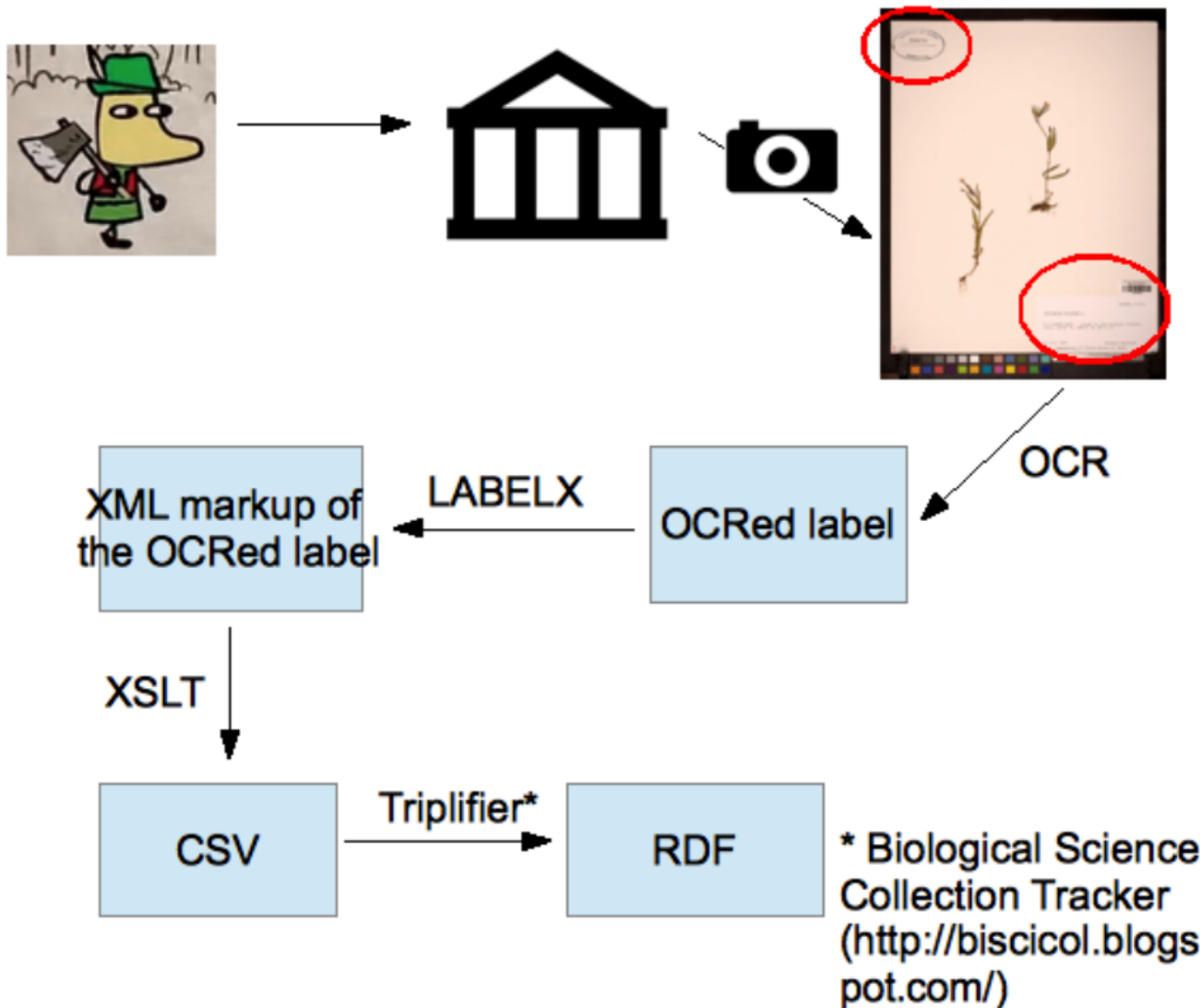




Figure 2.
Example Museum Specimen Label

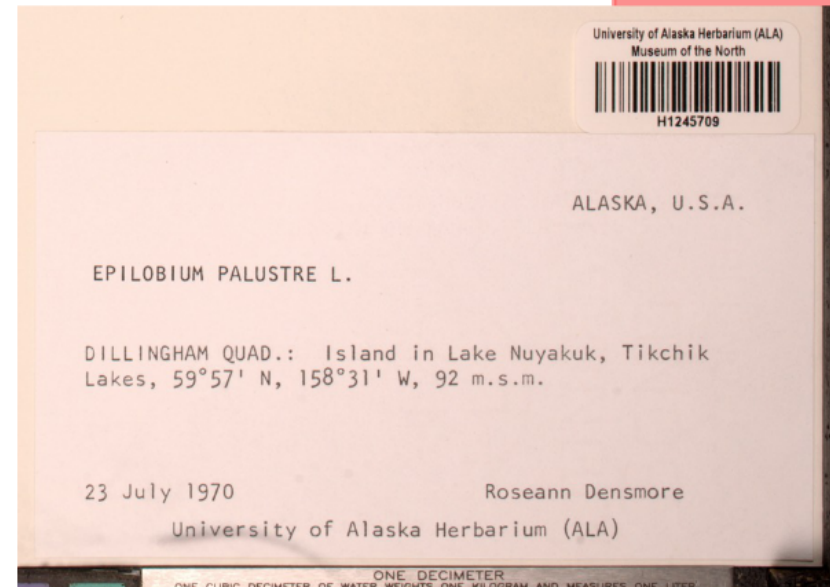
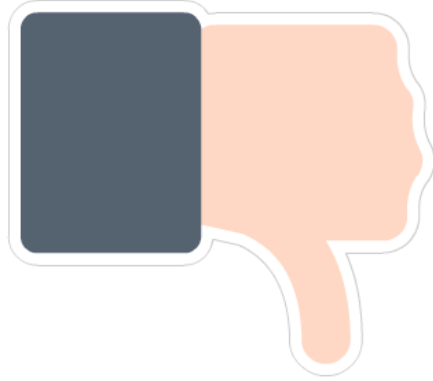


Figure 3. Example Label

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°57' N, 158°31' W, 92 m.s.m.
23 July 1970 Roseann Densmore
University of Alaska Herbarium (ALA)

Figure 4. OCRed label

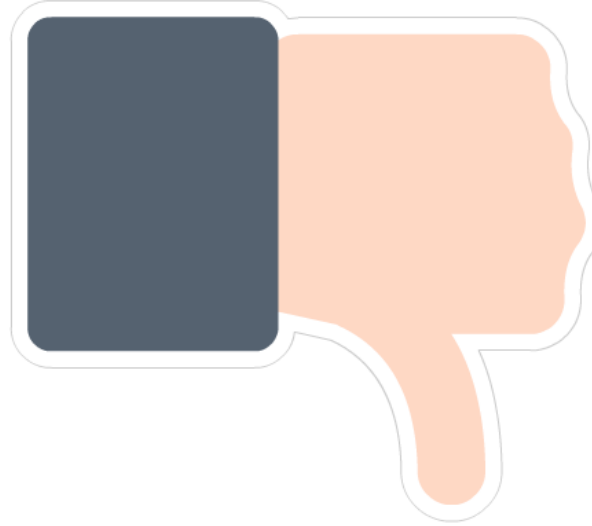


The Post Office did
this already.

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, 158°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

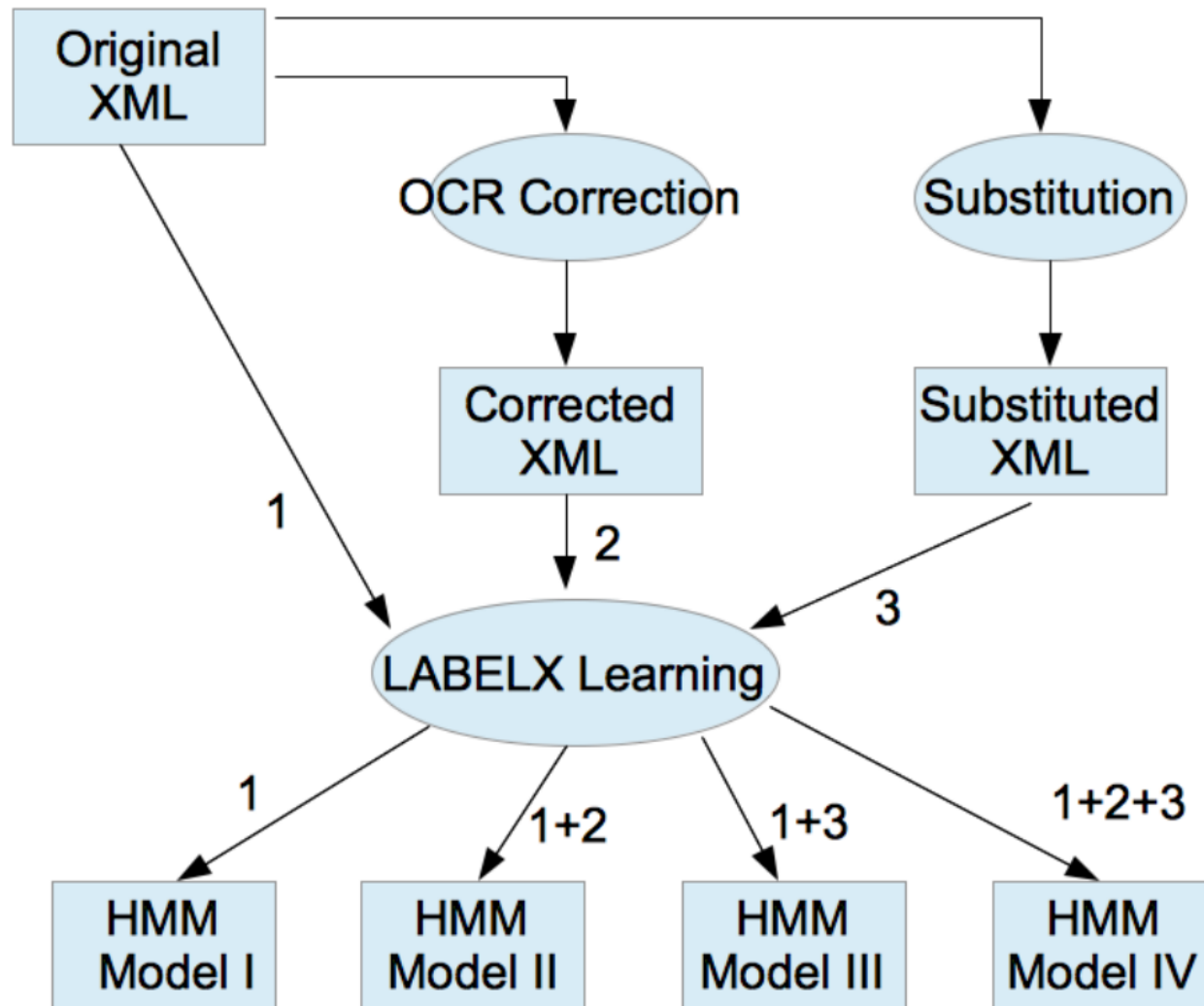


Figure 6. LABELX Learning



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

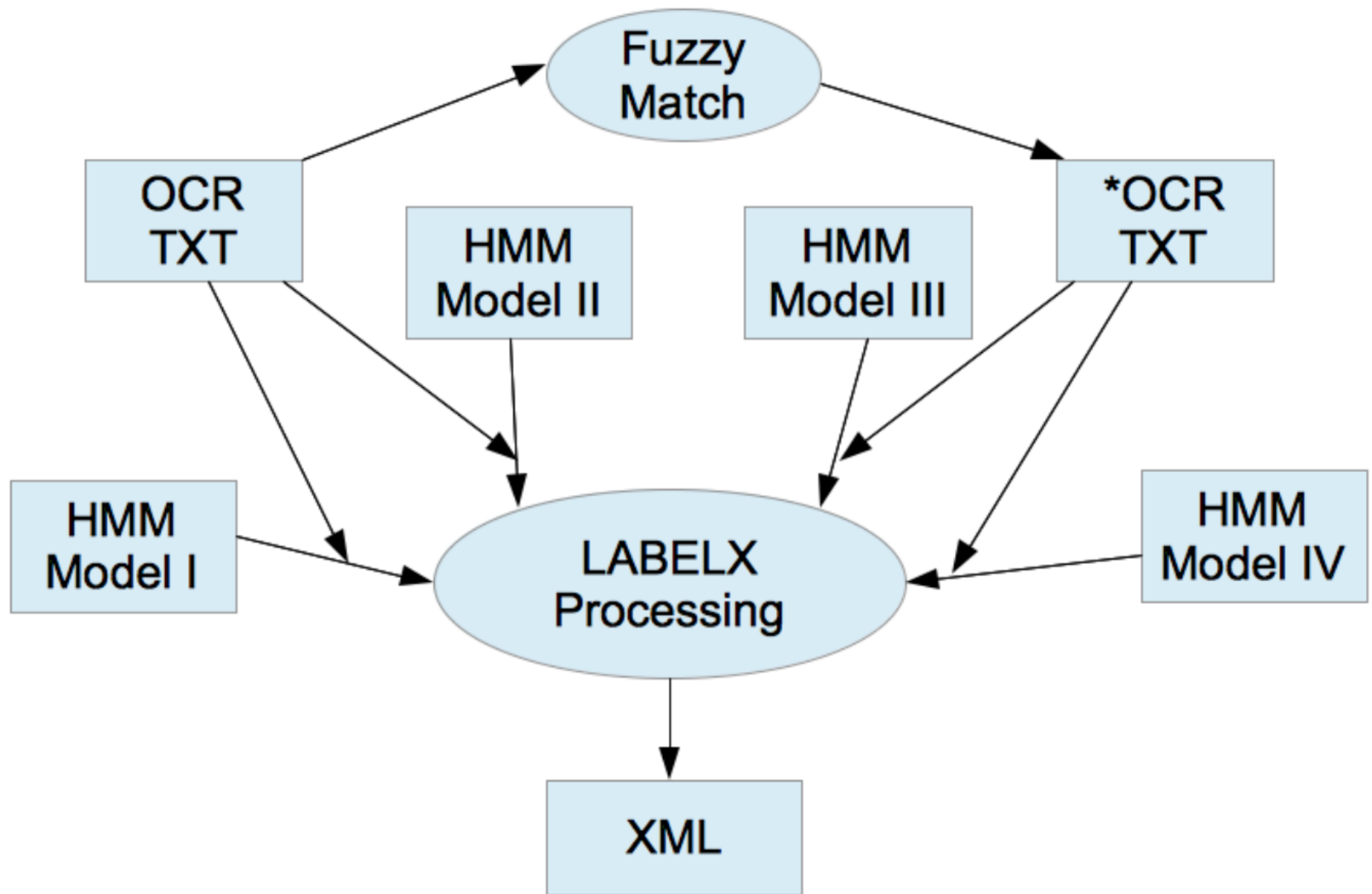


Figure 7. LABELX Parsing

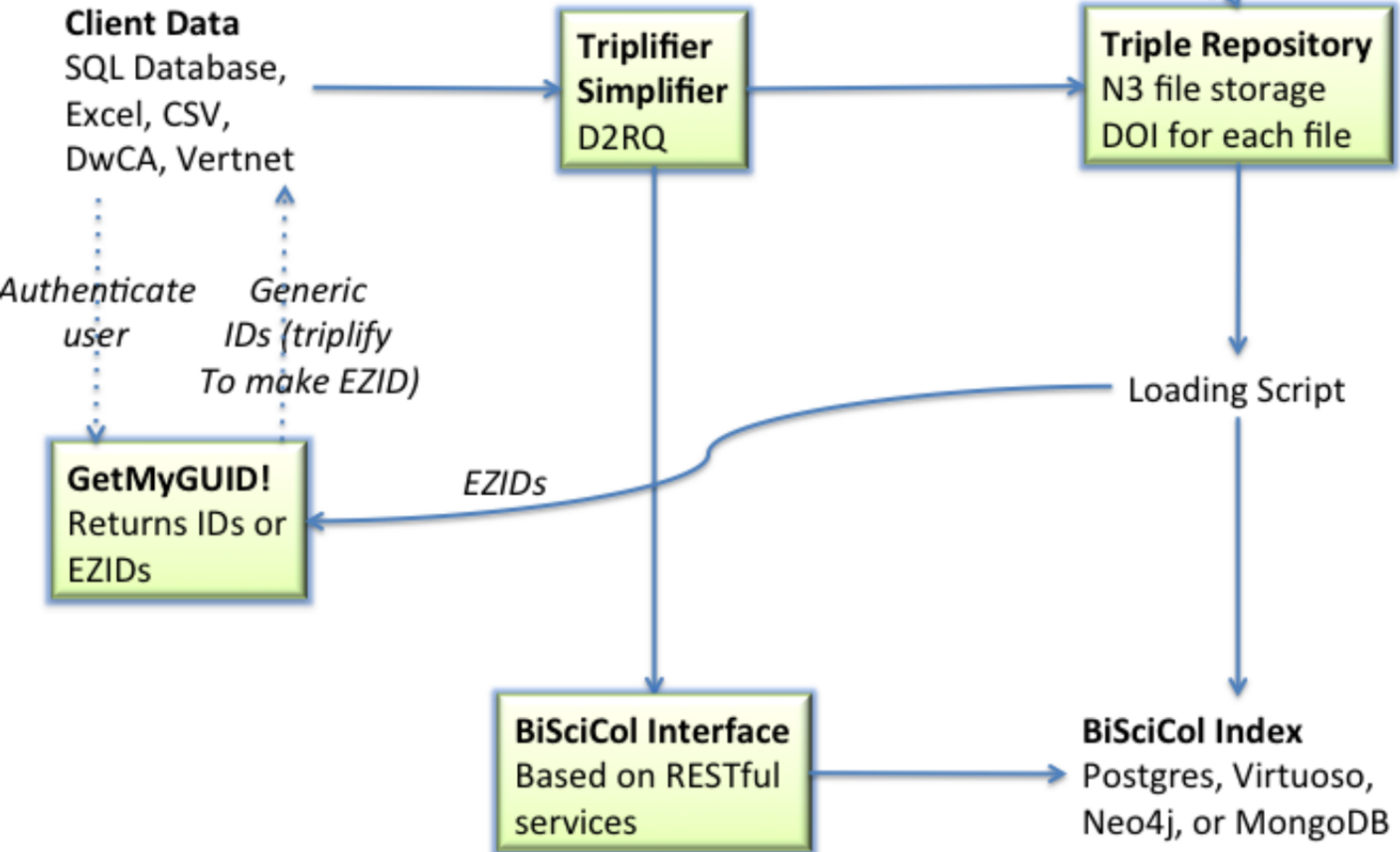
BiSciCol Relationship Terms

RDF	Definition	Example	Symmetric	Transitive
	Physical material (obj1) that is substantially derived from other physical material (obj2)	Tissue (obj1) derived from a Specimen (obj2)		Yes
	An entity (obj1) whose existence depends on another entity (obj2)	Specimen (obj1) part of a collecting event instance (obj2)		
	Two instances (obj1,obj2) that are understood to be the same thing	Obj1 and Obj2 both refer to the same specimen.	Yes	Yes
	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

BiSciCol
Services

Optional dump direct to N3 format



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, 158°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, 158°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

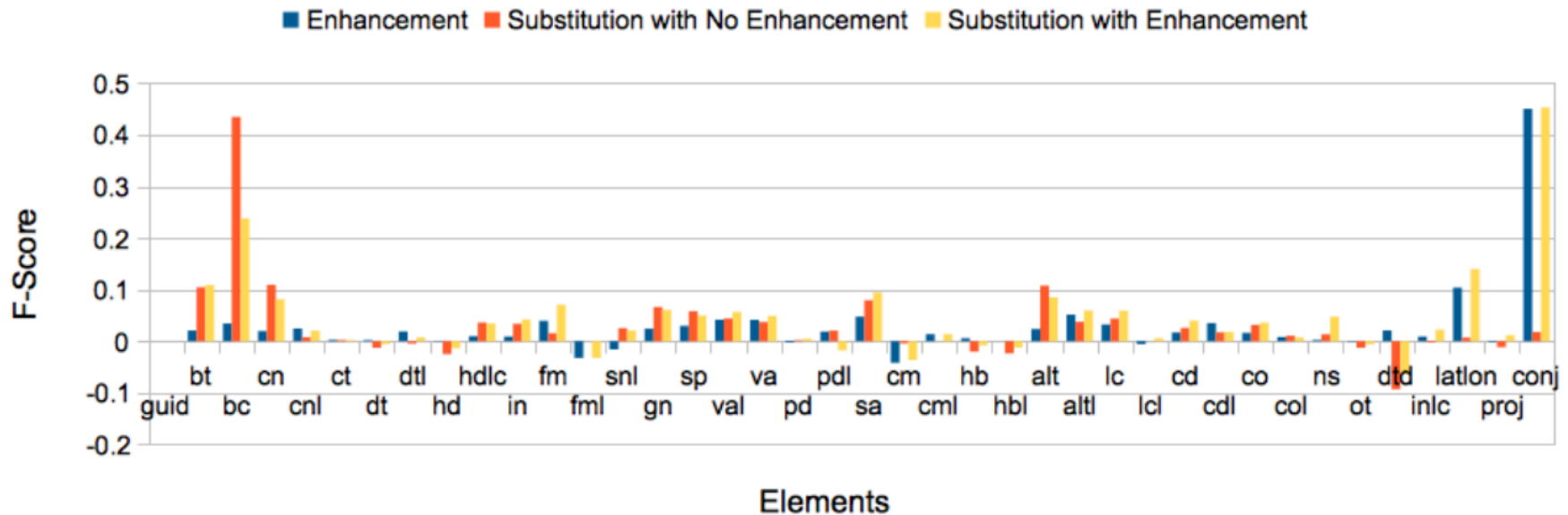


My technique works better on my 100
museum labels (collected between
2009-2011 in my backyard) on my
Commodor 64.

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.

Differences from the Control in F-Score





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	0.7584	+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

