Mapping Spreadsheets to RDF – Supporting Excel in RML

Markus Schröder, Christian Jilek, Andreas Dengel

Second International Workshop on Knowledge Graph Construction

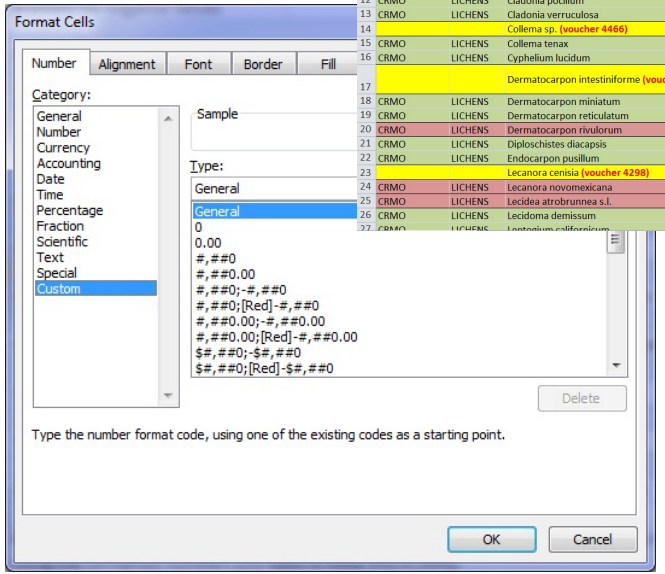
06.06.2021

A Spreadsheet

(or sometimes a canvas painted by a data artist)

- Spreadsheets
  - well understood
  - easy and fast possibility to enter data
  - complex workbooks
  - multiple sheets
  - cells having rich meta data
    - formats, colors, fonts, styles, borders, etc.
    - arbitrarily arranged
    - can lead to inconsistent and unstructured content

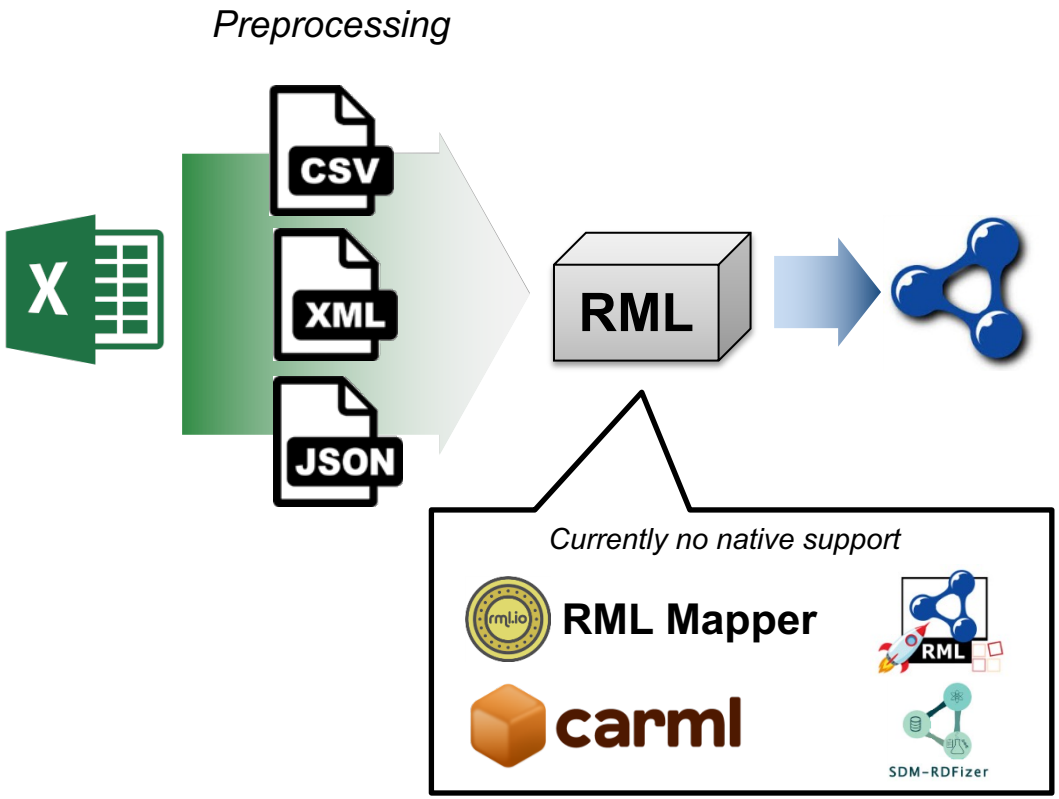
	A	B	C	D	E
1	Project_name	Class	Scientific_name	Scientific_name_Authority	Dfident
2	CRMO	LICHENS	Acarospora schleicheri	Acarospora schleicheri (Ach.) A. Massal.	-1
3	CRMO	LICHENS	Anaptychia elbursiana (Physconia thomsonii)	Anaptychia elbursiana (Szatala) Poelt	-1
4	CRMO	LICHENS	Arthonia glebosa	Arthonia glebosa Tuck.	0
5	CRMO	LICHENS	Caloplaca citrina	Caloplaca citrina (Hoffm.) Th. Fr.	-1
6			Caloplaca stillicidiorum (voucher 4229)	Caloplaca stillicidiorum (Vahl) Lynge (was not on this spreadsheet)	
7	CRMO	LICHENS	Candelaria concolor	Candelaria concolor (Dickson) Stein	-1
8			Candelaria reflexa (voucher 4299)	Candelaria reflexa (Nyl.) Lettau (was not on this spreadsheet)	?
9	CRMO	LICHENS	Cladonia chlorophaea	Cladonia chlorophaea (Flörke ex Sommerf.) Sprengel	-1
10	CRMO	LICHENS	Cladonia fimbriata	Cladonia fimbriata (L.) Fr.	-1
11	CRMO	LICHENS	Cladonia merochlorophaea	Cladonia merochlorophaea Asahina	-1
12	CRMO	LICHENS	Cladonia pocillum	Cladonia pocillum (Ach.) Grognot	-1
13	CRMO	LICHENS	Cladonia verruculosa	Cladonia verruculosa (Vainio) Ahti	-1
14			Collema sp. (voucher 4466)	Collema sp. F. H. Wigg. (was not on this spreadsheet)	
15	CRMO	LICHENS	Collema tenax	Collema tenax (Sw.) Ach.	-1
16	CRMO	LICHENS	Cyphelium lucidum	Cyphelium lucidum (Th. Fr.) Th. Fr.	-1
17			Dermatocarpon intestiniforme (voucher 4406)	Dermatocarpon intestiniforme (Korber) Hasse (was not on this spreadsheet)	
18	CRMO	LICHENS	Dermatocarpon miniatum	Dermatocarpon miniatum (L.) W. Mann	-1
19	CRMO	LICHENS	Dermatocarpon reticulatum	Dermatocarpon reticulatum H. Magn.	-1
20	CRMO	LICHENS	Dermatocarpon rivulorum	Dermatocarpon rivulorum (Arnold) Dalla Torre & Sarnth.	0
21	CRMO	LICHENS	Diploschistes diacapsis	Diploschistes diacapsis (Ach.) Lumbsch	-1
22	CRMO	LICHENS	Endocarpon pusillum	Endocarpon pusillum Hedwig	-1
23			Lecanora cenisia (voucher 4298)	Lecanora cenisia Ach. (was not on this spreadsheet)	
24	CRMO	LICHENS	Lecanora novomexicana	Lecanora novomexicana H. Magn.	0
25	CRMO	LICHENS	Lecidea atrobrunnea s.l.	Lecidea atrobrunnea (Ramond ex Lam. & DC.) Schaerer	0
26	CRMO	LICHENS	Lecidoma demissum	Lecidoma demissum (Rutstr.) Gotth. Schneider & Hertel	-1



Found at data.gov:
Hutten 2016_CRMO_lichen_moss_liverwort_.xlsx

From Spreadsheet to RDF

- Use input format that is supported
- Advantages of native support
  - Eliminates extraneous conversion efforts:
    No preprocessing and transformation needed
  - All aspects of a spreadsheet can be exploited
  - Eases mapping rule communication
  - For RML practitioners no extra language to learn



Related Work

Spread2RDF (Ruby syntax)

```
worksheet 'MaterialelementeKlassen',
          name:    :MaterialElementClasses,
          start:   :B5,
          subject: { uri: { namespace: PSM_MaterialElement },
                     type:         RDF::RDFS.Class,
                     sub_class_of: PSM.MaterialElement
          } do
  column :name, predicate: RDFS.label
  column :uri

  column :sub_class_of,     predicate: RDFS.subClassOf,
                            object:    { from: :MaterialElementClasses }
  column_block :parameter,  subject: { uri: :bnode, type: PSM.Parameter },
                            predicate: PSM.materialParameter,
                            statement: :restriction do
    column :name,        predicate: PSM.parameterName
    column :description, predicate: PSM.parameterDescription

    column :min,   predicate: PSM.parameterMinQuantity,
                   object:    { uri: :bnode, type: QUDT.QuantityValue },
                   &quantity_mapping
```

Mapping Master (Manchester syntax)

```
Class: @A*(rdfs:label 'analyte assay')
EquivalentTo:
(achieves_planned_objective some 'analyte measurement objective') and
(realizes some ('evaluant role' and (role_of some
                                        @D*(material_entity)))) and
(realizes some ('analyte role' and
                 (role_of some ('scattered molecular aggregate' and
                                 ('has grain' only
                                  @B*('molecular entity'))))))
SubClassOf:
 has_specified_output some
     ('scalar measurement datum' and
      ('is quality measurement of' some 'molecular concentration') and
       ('has measurement unit label' some
        @F*('measurement unit label')))
```

- Various languages with different features
- RML + Excel
  - Familiar syntax and concepts
  - Utilize RML features

Sheet2RDF (PEARL syntax)

```
prefix  :       <http://baseuri.org#>
prefix  coda:   <http://art.uniroma2.it/coda/contracts>
...

rule it.uniroma2.art.Sheet2RDFAnnotation id:row {
        nodes = {
                subject uri     col0_skos_Concept/value
                col1_skos_broader_value uri     col1_skos_broader/value
                col2_skos_prefLabel_value literal@en col2_skos_prefLabel/value
        }
        graph = {
                $subject rdf:type <http://www.w3.org/2004/02/skos/core#Concept> .
                $subject skos:inScheme <http://sheet2rdf#main> .
                OPTIONAL { $subject     skos:broader    $col1_skos_broader_value . }
                OPTIONAL { $subject     skos:prefLabel  $col2_skos_prefLabel_value . }
        }
}
```

XLWrap (TriG syntax)

```
{ [] a xl:Mapping ;
        ...
        xl:templateGraph :ANamedGraph
        ...
}

:ANamedGraph {
        [ xl:uri "'http://example.org/' & URLENCODE(A2 & B2)"^^xl:Expr ] a foaf:Person ;
        foaf:name           "A2 & ' ' & B2"^^xl:Expr ;
        foaf:mbox_sha1sum   "SHA(C2)"^^xl:Expr ;
}
```

Approach: Technical Integration in RMLMapper

- Extension to the RML Mapper tool
  [GitHub fork]

```
/** This class creates records based on RML rules ...3 lines */
public class RecordsFactory {

    private Map<Access, Map<String, Map<String, List<Record>>>> recordCache;
    private AccessFactory accessFactory;
    private Map<String, ReferenceFormulationRecordFactory> referenceFormulationRecordFactoryMap;

    public RecordsFactory(String basePath) {
        accessFactory = new AccessFactory(basePath);
        recordCache = new HashMap<>();

        referenceFormulationRecordFactoryMap = new HashMap<>();
        referenceFormulationRecordFactoryMap.put(NAMESPACES.QL + "XPath", new XMLRecordFactory());
        referenceFormulationRecordFactoryMap.put(NAMESPACES.QL + "JSONPath", new JSONRecordFactory());
        referenceFormulationRecordFactoryMap.put(NAMESPACES.QL + "CSV", new CSVRecordFactory());
        referenceFormulationRecordFactoryMap.put(NAMESPACES.QL + "Spreadsheet", new ExcelRecordFactory());
    }
```

```
RMLMapper [mschroeder-features]
  Source Packages
    be.ugent.rml
    be.ugent.rml.access
    be.ugent.rml.cli
    be.ugent.rml.conformer
    be.ugent.rml.extractor
    be.ugent.rml.functions
    be.ugent.rml.functions.lib
    be.ugent.rml.metadata
    be.ugent.rml.records
      CSVRecord.java
      CSVRecordFactory.java
      CSVW.java
      ExcelCell.java [-/M]
      ExcelRecord.java [-/M]
      ExcelRecordFactory.java [-/M]
      IteratorFormat.java
      JSONRecord.java
      JSONRecordFactory.java
```

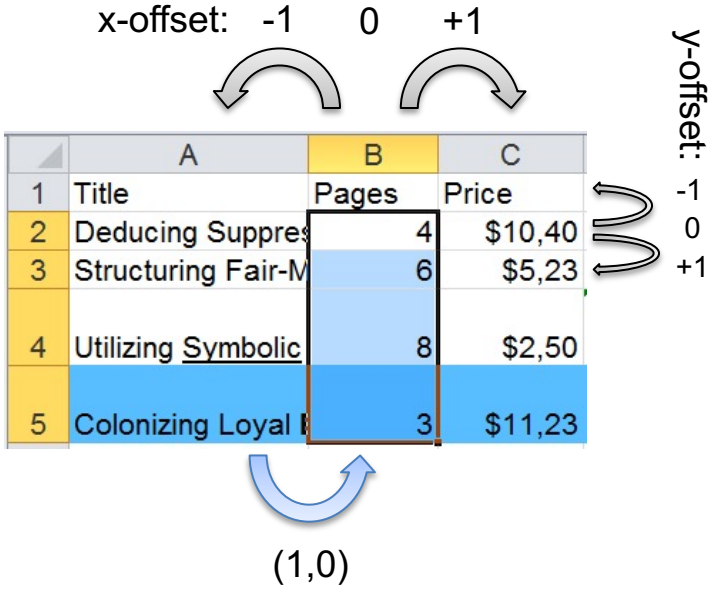
```
:ls1
    a rml:LogicalSource ;
    rml:referenceFormulation ql:Spreadsheet ;
    rml:source [
      a ss:Workbook;
      ss:url "workbook.xlsx" ;
      ss:sheetName "Papers" ;
      ss:range "A2:A5" ;
    ] .
```

- Record Factory
  - Use parameters to return records
  - APACHE POI
- Record
  - Interpret references

Conceptual Integration in RML

Cell Location Reference

- Cell iteration using ranges (e.g. B2:B5)
  - Instead of row iteration (like in CSV)
- Arbitrarily structured tables without any anchor points like column names
  - Relative reference to neighboring cells with parenthesis notation
  - Absolute reference to cells with square brackets

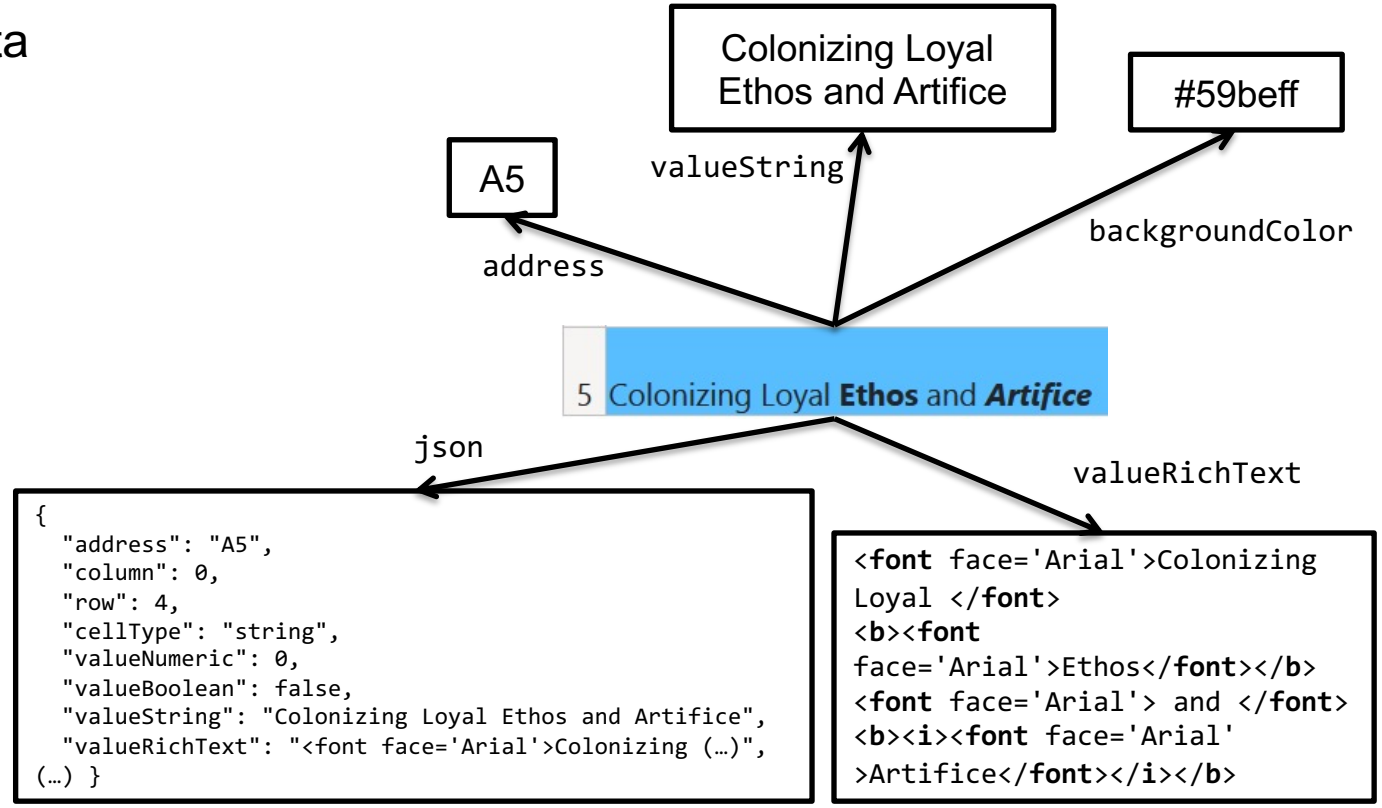



```
rr:predicateObjectMap [
  a rr:PredicateObjectMap ;
  rr:predicateMap  [
    a rr:PredicateMap ;
    rr:constant  ex:numberOfPages
  ] ;
  rr:objectMap [
    a rr:ObjectMap ;
    rml:reference  "(1,0).valueInt" ;
    rr:datatype xsd:integer
  ]
] .
```

Conceptual Integration in RML

Meta Data References

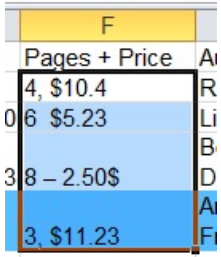
- Access cell meta data
  - address
  - column
  - row
  - backgroundColor
  - foregroundColor
  - fontColor
  - fontName
  - fontSize
  - valueNumeric
  - valueInt
  - valueBoolean
  - valueFormula
  - valueError
  - valueString
  - valueRichText
  - json
  - value



Experimental Features
Multiple Different Properties in a Cell

- Each piece of information corresponds to a different property
- Zip together
  - properties
  - returned objects from FnO function
- Just a shortcut
  - Usually requires separate predicate object maps

```
rr:predicateObjectMap [
  a rr:PredicateObjectMap ;
  ss:zip true ;
  rr:predicateMap  [
    a rr:PredicateMap ;
    rr:constant ( ex:numberOfPages ex:price )
  ] ;
  rr:objectMap [
      a rr:ObjectMap , fnml:FunctionMap ;
      fnml:functionValue  [
          rr:predicateObjectMap  [
              rr:predicate  fno:executes ;
              rr:object     <java:ifRegexReturnGroup>
          ] ;
          # String value
          rr:predicateObjectMap [
              rr:predicate  <java:parameter.predicate.string.0> ;
              rr:objectMap  [ rml:reference  "(5,0).valueString" ]
          ] ;
          # String pattern
          rr:predicateObjectMap [
              rr:predicate  <java:parameter.predicate.string.1> ;
              rr:objectMap  [ rr:constant  "\\d+\\.?\\d*" ]
          ] ;
```



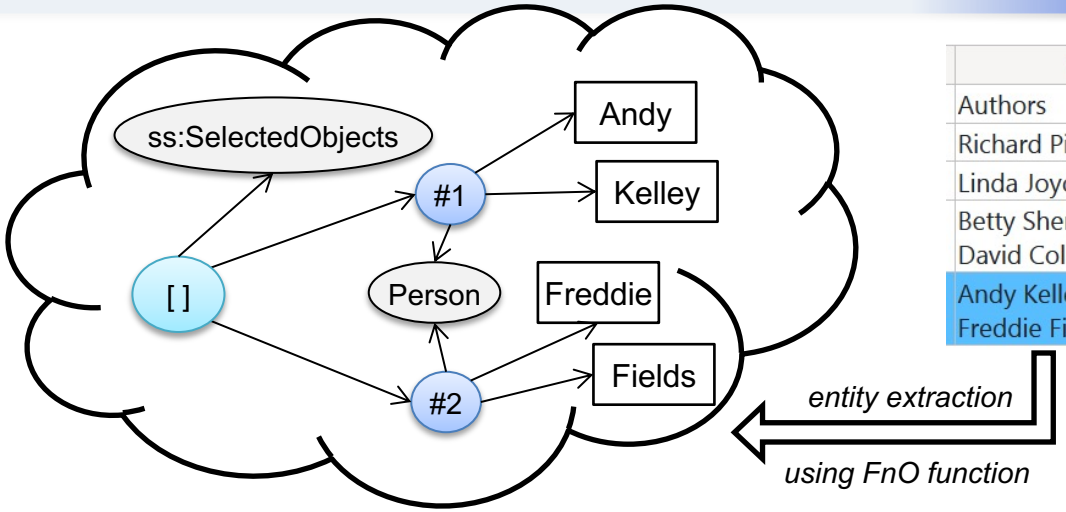
page information *price information*

```
<http://example.org/A2> a <http://example.org/Paper>;
  <http://example.org/numberOfPages> 4.0;
  <http://example.org/price> 10.4 .

<http://example.org/A3> a <http://example.org/Paper>;
  <http://example.org/numberOfPages> 6.0;
  <http://example.org/price> 5.23 .

<http://example.org/A4> a <http://example.org/Paper>;
  <http://example.org/numberOfPages> 8.0;
  <http://example.org/price> 2.5 .

<http://example.org/A5> a <http://example.org/Paper>;
  <http://example.org/numberOfPages> 3.0;
  <http://example.org/price> 11.23 .
```

Result

Experimental Features
Multiple Complex Entities in a Cell

- For example: list of persons having first and last names
- RDF graph in turtle syntax
- New term type `ss:Graph`
- Graph is added to result
- Selected ones are mapped using `ss:SelectedObjects`



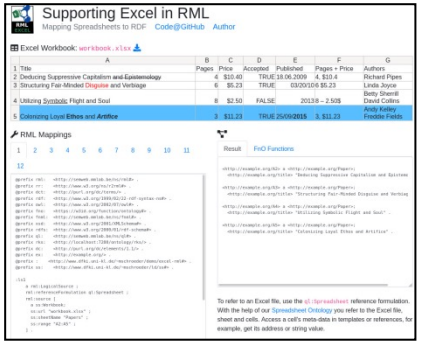
G
Authors
Richard Pipes
Linda Joyce
Betty Sherrill David Collins
Andy Kelley Freddie Fields

```
<http://example.org/A5> a <http://example.org/Paper>;
  <http://example.org/hasAuthor>
    <uuid:5107e5c6-48d2-4981-9965-890716cde15c>,
    <uuid:2d1fd69b-a742-475e-b382-5e2f7d30fa07> .

<uuid:5107e5c6-48d2-4981-9965-890716cde15c> a <http://xmlns.com/foaf/0.1/Person>;
  <http://xmlns.com/foaf/0.1/firstName> "Andy";
  <http://xmlns.com/foaf/0.1/lastName> "Kelley" .

<uuid:2d1fd69b-a742-475e-b382-5e2f7d30fa07> a <http://xmlns.com/foaf/0.1/Person>;
  <http://xmlns.com/foaf/0.1/firstName> "Freddie";
  <http://xmlns.com/foaf/0.1/lastName> "Fields" .
```

Result

Conclusion

- Implemented Excel support in RML Mapper
- Cell iteration, location, meta data access
- Try it out on our demo page
  - GitHub code



- You may also be interested in
  - Data Sprout – Dataset generation for evaluating KGC

- More related research in our project
  - SensAI

Thank you for your attention.
Questions?