ontotext

Large-scale Reasoning with a Complex Cultural Heritage Ontology (CIDOC CRM)

Vladimir Alexiev, Dimitar Manov, Jana Parvanova, Svetoslav Petrov

Practical Experiences with CIDOC CRM and its Extensions (CRMEX 2013)

TPDL 2012, 26 Sep 2013, Malta

Agenda

- ResearchSpace project
- RS Semantic Search
- Fundamental Relation (FR) search
- Implemented FRs
- OWLIM Rules
- Example: FR92i_created_by
- Sub-FRs and Dependency Graph
- Complexity: Classes (Type statements)
- Complexity: Properties
- Comparison to Other Repositories
- Performance of Straight SPARQL Implementation
- Performance of OurImplementation

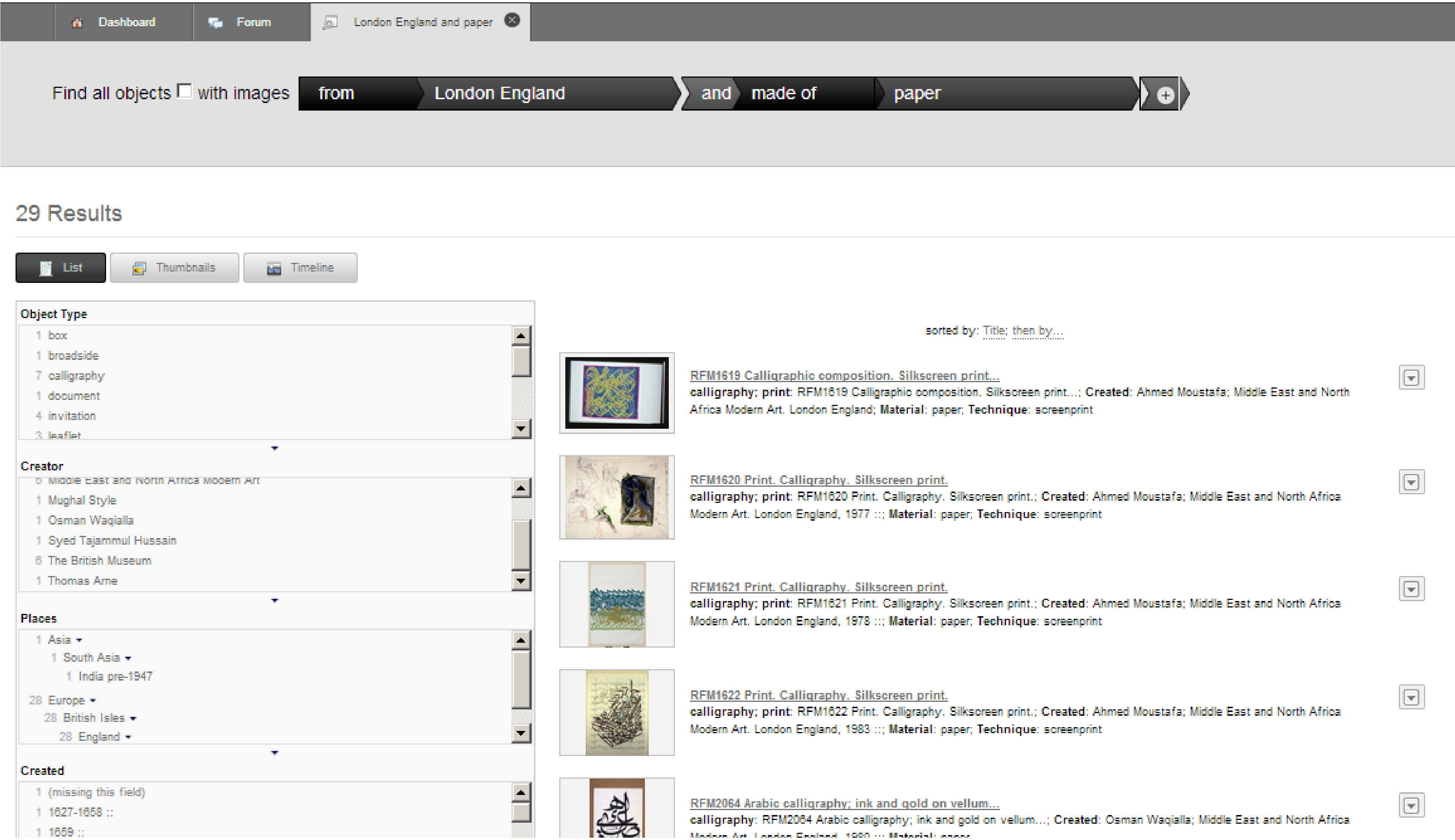
ResearchSpace (RS) Project

- Funded by Mellon Foundation, run by the British Museum, sw dev by Ontotext
  - Stage 3 (Working Prototype): developed between Nov 2011 and Apr 2013.
  - Stage 4: expected to start in 2013, with more development and more museums/galleries on board
- Support collaborative research projects for CH scholars
  - Open source framework and hosted environment for web-based research, knowledge sharing and web publishing
- Intends to provide:
  - Data conversion and aggregation (LIDO/CDWA/similar to CIDOC CRM)
  - Semantic search based on Fundamental Relations
  - Collaboration tools, such as forums, tags, data baskets, sharing, dashboards
  - Research tools , such as Image Annotation, Image Compare, Timeline and Geographical Mapping...
  - Web Publication
- Semantic technology is at the core of RS because it provides effective data integration across different organizations and projects.
  - Uses Ontotext's OWLIM repository: powerful reasoning (equivalent to OWL2 RL), fast performance, efficient multi-user access, full SPARQL 1.1 support, incremental assert and retract

RS Semantic Search

- Allows a user that is not familiar with CRM or the BM data to perform simple and intuitive searches.
- Features:
  - Intuitive "sentence-based" UI
  - Searches can be saved, bookmarked (put in a "data basket"), edited, shared between users
  - Auto-completion across all searchable thesauri. Available search relations and appropriate Thesauri are coordinated
  - Search across datasets. E.g. once the entity "Rembrandt" is co-referenced between the BM People and RKD Artists thesauri, paintings by Rembrandt can be found across the BM and RKD datasets
  - Faceting of search results
  - Details, thumbnails (lightbox), list, timeline view
  - Put search result to data basket, invoke RS tool

RS Search: Example 1



Dashboard | Forum | London England and paper

Find all objects □ with images **from** London England **and** made of paper

29 Results

List | Thumbnails | Timeline

Object Type
1 box
1 broadside
7 calligraphy
1 document
4 invitation
3 leaflet

Creator
6 Middle East and North Africa Modern Art
1 Mughal Style
1 Osman Waqialla
1 Syed Tajammul Hussain
6 The British Museum
1 Thomas Arne

Places
1 Asia
1 South Asia
1 India pre-1947
28 Europe
28 British Isles
28 England

Created
1 (missing this field)
1 1627-1658 ::
1 1659 ::

sorted by: Title; then by...

RFM1619 Calligraphic composition. Silkscreen print...
calligraphy; print: RFM1619 Calligraphic composition. Silkscreen print...; **Created**: Ahmed Moustafa; Middle East and North Africa Modern Art. London England; **Material**: paper; **Technique**: screenprint

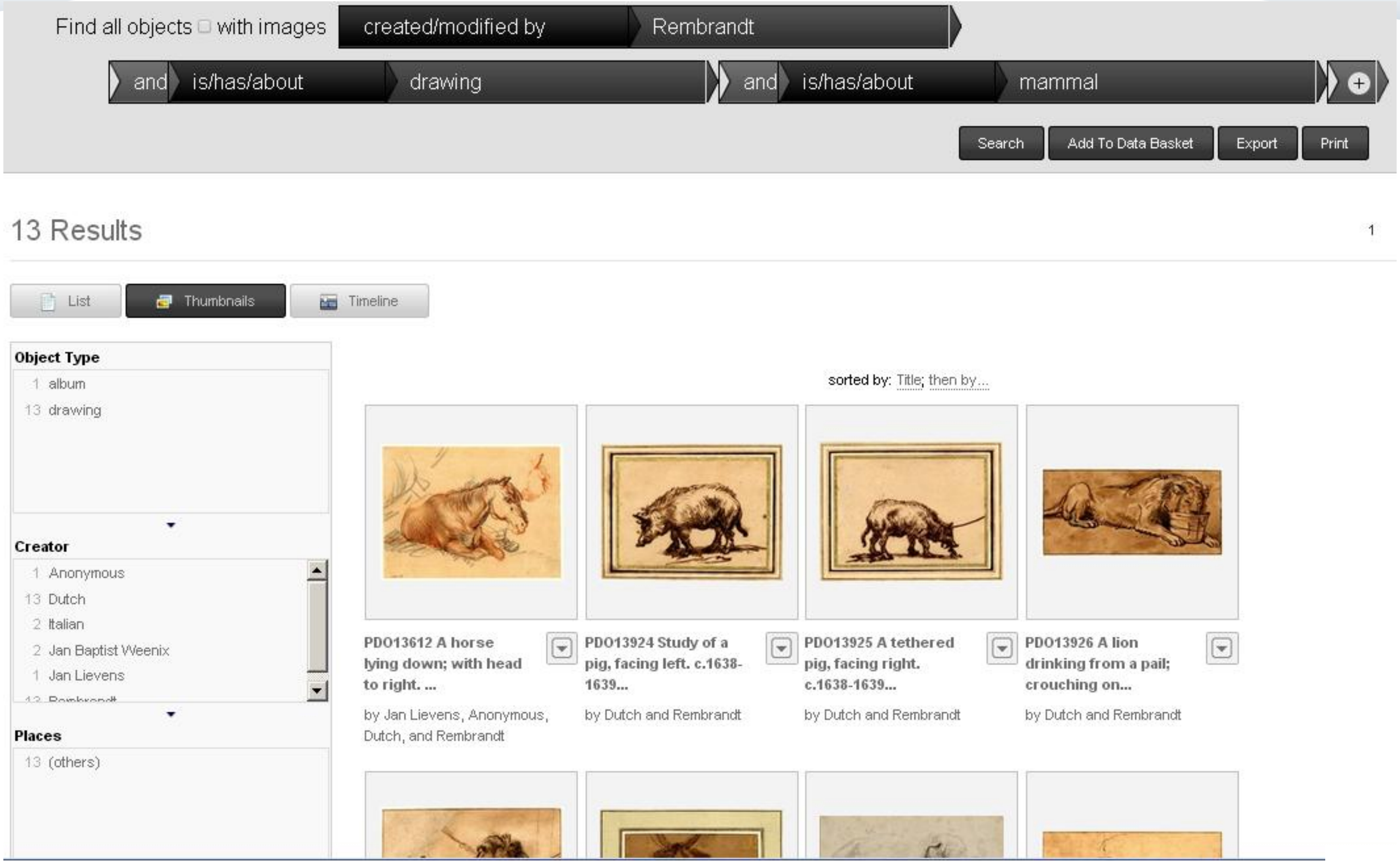
RFM1620 Print. Calligraphy. Silkscreen print.
calligraphy; print: RFM1620 Print. Calligraphy. Silkscreen print.; **Created**: Ahmed Moustafa; Middle East and North Africa Modern Art. London England, 1977 :::; **Material**: paper; **Technique**: screenprint

RFM1621 Print. Calligraphy. Silkscreen print.
calligraphy; print: RFM1621 Print. Calligraphy. Silkscreen print.; **Created**: Ahmed Moustafa; Middle East and North Africa Modern Art. London England, 1978 :::; **Material**: paper; **Technique**: screenprint

RFM1622 Print. Calligraphy. Silkscreen print.
calligraphy; print: RFM1622 Print. Calligraphy. Silkscreen print.; **Created**: Ahmed Moustafa; Middle East and North Africa Modern Art. London England, 1983 :::; **Material**: paper; **Technique**: screenprint

RFM2064 Arabic calligraphy; ink and gold on vellum...
calligraphy: RFM2064 Arabic calligraphy; ink and gold on vellum...; **Created**: Osman Waqialla; Middle East and North Africa Modern Art. London England, 1980 ::: **Material**: paper

RS Search: Example 2



Find all objects □ with images | created/modified by | Rembrandt

and | is/has/about | drawing | and | is/has/about | mammal | ⊕

Search | Add To Data Basket | Export | Print

13 Results

1

List | Thumbnails | Timeline

Object Type

1 album
13 drawing

Creator

1 Anonymous
13 Dutch
2 Italian
2 Jan Baptist Weenix
1 Jan Lievens
13 Rembrandt

Places

13 (others)

sorted by: Title; then by...

PD013612 A horse lying down; with head to right. ...
by Jan Lievens, Anonymous, Dutch, and Rembrandt

PD013924 Study of a pig, facing left. c.1638-1639...
by Dutch and Rembrandt

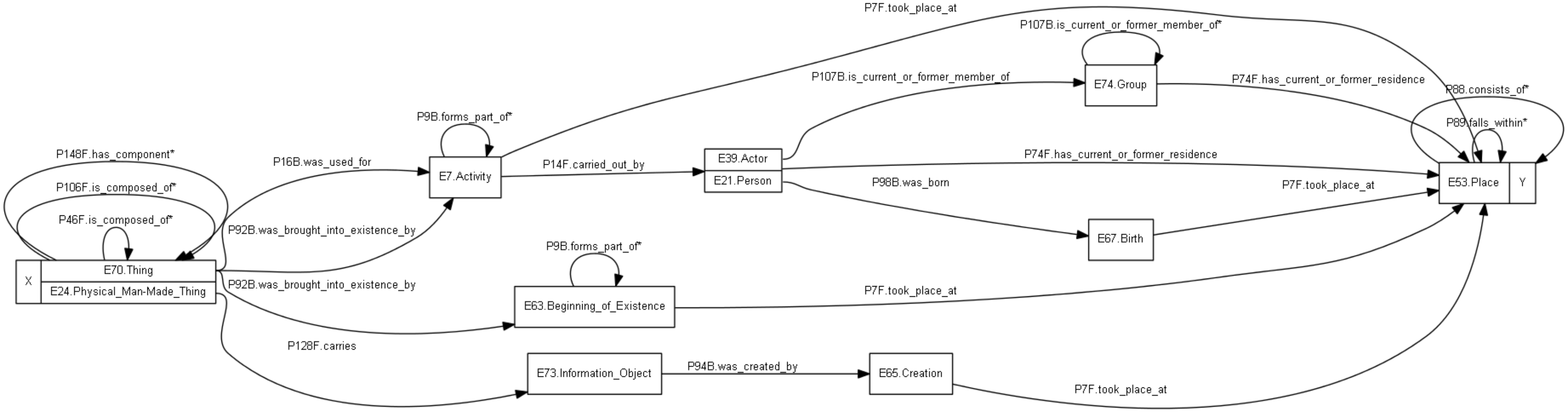
PD013925 A tethered pig, facing right. c.1638-1639...
by Dutch and Rembrandt

PD013926 A lion drinking from a pail; crouching on...
by Dutch and Rembrandt

- Finds narrower terms
- RS Video by Dominic Oldman (RS PI and BM IT dev manager) http://www.youtube.com/watch?v=HCnwgq6ebAs

Fundamental Relation (FR) Search

- How does a user search through a large CRM network?
- An answer: Fundamental Relations.
  - Aggregate a large number of paths through CRM data into a smaller number of searchable relations.
  - Provide a "search index" over the CRM relations
- E.g.: FR "Thing from Place"



- Initial implementation presented at SDA 2012 (TPDL 2012), Sep 2012, Cyprus (CEUR WS Vol.912)

Implemented FRs

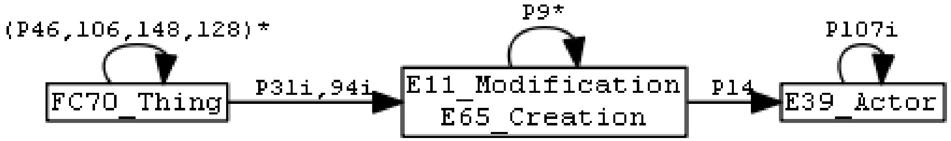
N	FR	Description
1	FR92i_created_by	Thing (or part/inscription thereof) was created or modified/repaired by Actor (or group it is member of, e.g. Nationality)
2	FR15_influenced_by	Thing's production was influ-enced/motivated by Actor (or group it is member of). E.g.: Manner/ School/ Style of; or Issuer, Ruler, Magistrate who authorised, patronised, ordered the produc-tion.
3	FR52_current_owner_keeper	Thing has current owner or keeper Actor
4	FR51_former_or_current_owner_keeper	Thing has former or current owner or keeper Actor, or ownership/custody was transferred from/to actor in Acquisition/Transfer of Custody event
5	FR67_about_actor	Thing depicts or refers to Actor, or carries an information object that is about Actor, or bears similarity with a thing that is about Actor
6	FR12_has_met	Thing (or another thing it is part of) has met actor in the same event (or event that is part of it)
7	FR67_about_period	Thing depicts or refers to Event/Period, or carries an information object that is about Event, or bears similarity with a thing that is about Event
8	FR12_was_present_at	Thing was present at Event (eg exhi-bition) or is from Period
9	FR92i_created_in	Thing (or part/inscription thereof) created or modified/repaired at/in place (or a broader containing place)
10	FR55_located_in	Thing has current or permanent location in Place (or a broader containing place)
11	FR12_found_at	Thing was found (discovered, excavated) at Place (or a broader containing place)
12	FR7_from_place	Thing has former, current or permanent location at place, or was created/found at place, or moved to/from place, or changed ownership/custody at place (or a broader containing place)
13	FR67_about_place	Thing depicts or refers to a place or fea-ture located in place, or is similar in features or composed of or carries an infor-mation object that depicts or refers to a place
14	FR2_has_type	Thing is of Type, or has Shape, or is of Kind, or is about or depicts a type (e.g. IconClass or subject heading)
15	FR45_is_made_of	Thing (or part thereof) consists of ma-terial
16	FR32_used_technique	The production of Thing (or part thereof) used general technique
17	luc:myIndex	The full text of the thing's description (including the-saurus terms and textual descriptions) matches the given keyword. FTS using Lucene built into OWLIM.
18	FR108i_82_produced_within	Thing was created within an interval that intersects the given interval or year.
19	FR1_identified_by	Thing (or part thereof) has Identifier. Exact-match string
20	FR138i_has_representation	Thing has at least one image repre-sentation. Used to select objects that have images
21	FR138i_representation	Thing has image representation. Used to fetch all images of an object
22	FR_main_representation	Thing has main image representation. Used to display object thumbnail in search results
23	FR_dataset	Thing belongs to indicated dataset. Used for faceting by dataset

OWLIM Rules

- OWLIM reasoning features:
  - Custom rule-sets. The standard semantics that OWLIM supports (RDFS, RDFS Horst, OWL RL, QL and DL) are also implemented as rulesets.
  - Fully-materializing forward-chaining reasoning. Rule consequences are stored in the repository and query answering is very fast.
  - sameAs optimization that allows fast cross-collection search using coreferenced values
  - Incremental retraction: when a triple is deleted, OWLIM removes all inferred consequences that are left without support (recursively)
  - Incremental insert: when a triple is inserted (even an ontology triple), all rules are checked. If a rule fires, the new conclusion is also checked against the rules, etc.
  - Efficient rule execution: rules are compiled to Java and executed quickly
- **120 OWLIM Rules** to implement 23 FRs:
  - 14 rules implement RDFS reasoning, owl:TransitiveProperty, owl:inverseOf (OWL) and ptop:transitiveOver (PROTON)
  - 106 rules implement FRs. Used a method of decomposing an FR to sub-FR : conjunctive (e.g. checking the type of a node), disjunctive (parallel), serial (property path), transitive

Example: FR92i_created_by

- Thing created by Actor
  - Thing (or part/inscription thereof) was created or modified/repaired by Actor (or a group it is a member of)

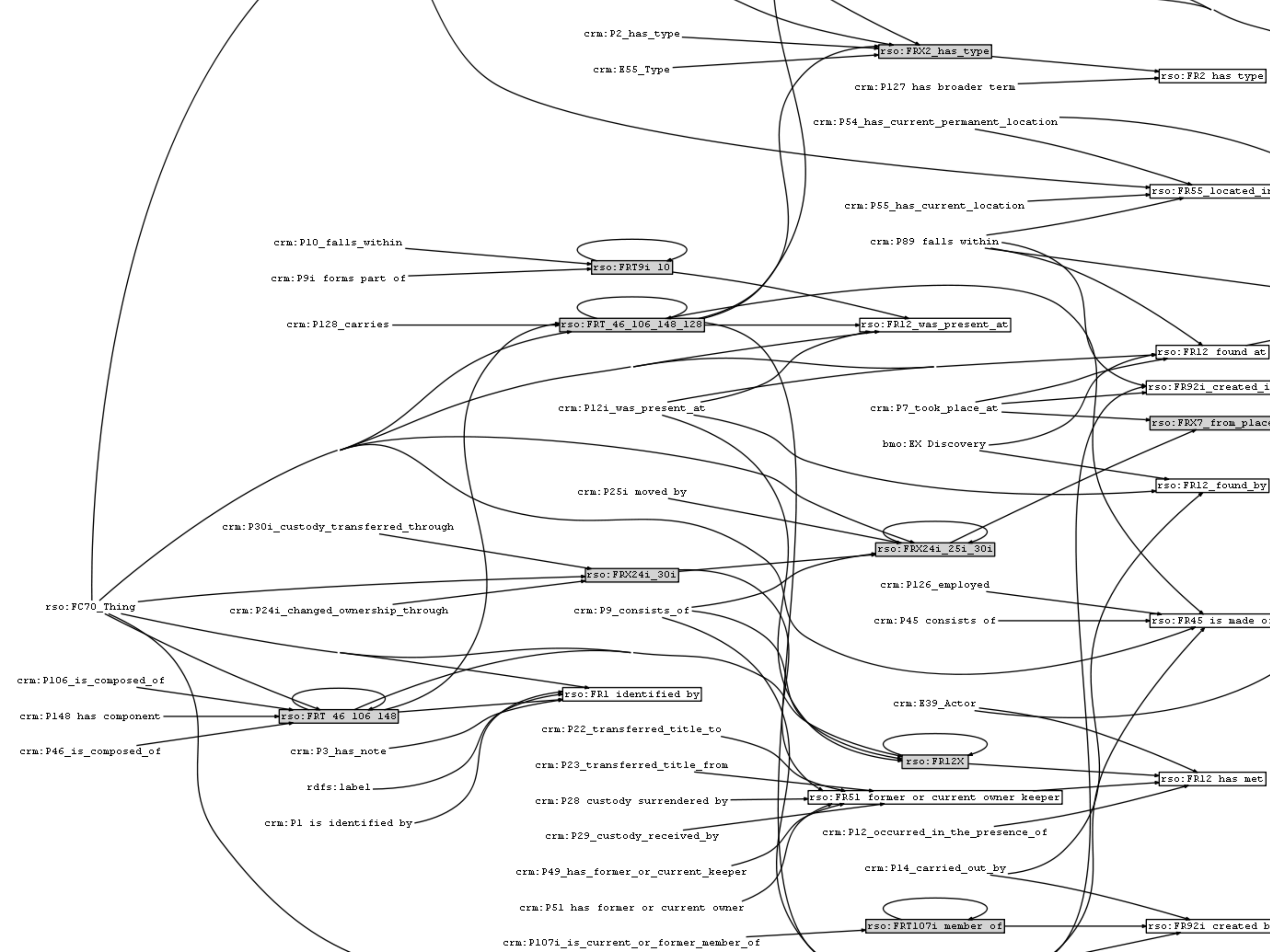


- Source properties:
  - P46_is_composed_of, P106_is_composed_of, P148_has_component: navigates object part hierarchy
  - P128_carries: to transition from object to Inscription carried by it
  - P31i_was_modified_by (includes P108i_was_produced_by), P94i_was_created_ by: Modification/Production of physical thing, Creation of conceptual thing (Inscription)
  - P9_consists_of: navigates event part hierarchy (BM models uncorrelated production facts as sub-events)
  - P14_carried_out_by, P107i_is_current_or_former_member_of: agent and groups he's member of
- Sub-FRs
  - FRT_46_106_148_128 := (P46|P106|P148|P128)+
  - FRX92i_created := (FC70_Thing) FRT_46_106_148_128* / (P31i | P94i) / P9*
  - FR92i_created_by := FRX92i_created / P14 / P107i*

Rules for FR92i_created_by

- Use a simple shortcut notation
  - Script translates ";" to newline and "=>" to "---------"
  - Also weaves from wiki
  - Checks variable linearity
  - Generates dependency graph (see next)
- 10 rules for FRT_46_106_148_128
- 7 rules for FR92i_created_by:

```
x <rdf:type> <rso:FC70_Thing>; x <crm:P31i_was_modified_by> y => x <rso:FRX92i_created> y
x <rdf:type> <rso:FC70_Thing>; x <crm:P94i_was_created_by>  y => x <rso:FRX92i_created> y
x <rso:FRT_46_106_148_128> y; y <crm:P31i_was_modified_by> z => x <rso:FRX92i_created> z
x <rso:FRT_46_106_148_128> y; y <crm:P94i_was_created_by>  z => x <rso:FRX92i_created> z
x <rso:FRX92i_created> y; y <crm:P9_consists_of> z => x <rso:FRX92i_created> z
x <rso:FRX92i_created> y; y <crm:P14_carried_out_by> z => x <rso:FR92i_created_by> z
x <rso:FRX92i_created> y; y <crm:P14_carried_out_by> z; z <rso:FRT107i_member_of> t
    => x <rso:FR92i_created_by> t
```


crm:P2_has_type
rso:FRX2_has_type
crm:E55_Type
rso:FR2 has type
crm:P127 has broader term
crm:P54_has_current_permanent_location
rso:FR55_located_i
crm:P55_has_current_location
crm:P89 falls within
crm:P10_falls_within
rso:FRT9i 10
crm:P9i forms part of
crm:P128_carries
rso:FRT_46_106_148_128
rso:FR12_was_present_at
rso:FR12 found at
rso:FR92i_created_i
crm:P12i_was_present_at
crm:P7_took_place_at
rso:FRX7_from_place
bmo:EX Discovery
rso:FR12_found_by
crm:P25i moved by
crm:P30i_custody_transferred_through
rso:FRX24i_25i_30i
rso:FRX24i_30i
crm:P126_employed
rso:FC70_Thing
crm:P24i_changed_ownership_through
crm:P9_consists_of
crm:P45 consists of
rso:FR45 is made o
crm:P106_is_composed_of
rso:FR1 identified by
crm:E39_Actor
crm:P148 has component
rso:FRT 46 106 148
crm:P22_transferred_title_to
rso:FR12X
crm:P46_is_composed_of
crm:P3_has_note
crm:P23_transferred_title_from
rso:FR12 has met
rdfs:label
crm:P28 custody surrendered by
rso:FR51 former or current owner keeper
crm:P1 is identified by
crm:P29_custody_received_by
crm:P12_occurred_in_the_presence_of
crm:P49_has_former_or_current_keeper
crm:P14_carried_out_by
crm:P51 has former or current owner
rso:FRT107i member of
rso:FR92i created b
crm:P107i_is_current_or_former_member_of

Volumetrics

- Museum objects: 2,051,797 (most from the British Museum)
  - Currently completing the ingest of Yale Center for British Art objects to RS (50k)
- Thesaurus entries: 415,509 (skos:Concept)
  - All kinds of "fixed" values that are used for search: object types, materials, techniques, people, places, … (a total of 90 ConceptSchemes)
- Explicit statements: 195,208,156. We estimate that of these:
  - 185M are for objects (90 statements/object)
  - 9M are for thesaurus entries (22 statements/term)
- Total statements: 916,735,486.
  - Expansion ratio is 4.7x (i.e. for each statement, 3.7 more are inferred)
  - Considerably higher compared to the typical expansion for general datasets
- Nodes (unique URLs and literals): 53,803,189 (don't use blank nodes)
- Repository size: 42 Gb
  - Object full-text index: 2.5 Gb, thesaurus full-text index (used for search auto-complete): 22Mb.
- Loading time (including all inferencing):
  - 22.2h on RAM drive
  - 32.9h on hard-disks

Complexity: Classes (Type statements)

Class	**Statement**
owl:Thing	36485904
E1_CRM_Entity	36485903
E77_Persistent_Item	17408450
E70_Thing	17339714
E71_Man-Made_Thing	17216212
E72_Legal_Object	17192518
E28_Conceptual_Object	14776488
E90_Symbolic_Object	14629292
E2_Temporal_Entity	11924877
E4_Period	11924877
E5_Event	11922986
E7_Activity	11796470
E63_Beginning_of_Existence	6377421
E11_Modification	6296015
E12_Production	6295825
rso:FC70_Thing	2051797
skos:Concept	415509
Total	**302149587**

Lawyers of the world, rejoice!

museum objects

Terms, people, places, materials, techniques..

- 238 classes, some of the top are summarizes in the table
- 415k skos:Concept (terms)
- 2M FC70_Thing (museum objects)
- Hierarchy is 10 levels deep : E1>E77>E70>E71>E28>E90> E73>E36>E37>E34
- For each Inscription, 12 type statements are inferred
- 6.3M E12_Production, repeated as the super-class E11_Modification, plus a few hundred Repairs
- Each E12 also repeated as E63_Beginning_of_Existence; plus 100k Birth and Formation
- Each E7 repeated as E5_Event, which is repeated as E4_Period (plus 19k historic Periods) and E2_Temporal_Entity
- ➢ 37% of all statements are type statements!

Reduce types: owl:Restriction vs Mixins

- Erlangen CRM states owl:Restrictions, e.g.:

  E72_Legal_Object SubClassOf: E70_Thing,
  P104_is_subject_to some E30_Right,
  P105_right_held_by some E39_Actor

  - M.Doerr has criticized this for ontological over-commitment
  - We don't need them so we cut them with XQuery tool deriving simpler profiles
- E72_Legal_Object:
  - Scope note: "material or immaterial items to which instances of E30 Right, such as the right of ownership or use, **can** be applied"
  - Do we really need it it in the main hierarchy?
- Just state P104 domain, and E72 will be inferred as needed
  - Akin to Common Lisp **mixins** or Ruby **traits**
- PSNC gives up rdfs:subClassOf inference
  - Using OWLIM custom rules (flexibility is good!)
  - For one node, all classes can be found with SPARQL 1.1 Path queries
  - May be a bit drastic…

Complexity: Properties

Properties	Statements	Percent
rdf:type	302149587	37.50%
Objects: CRM, rdfs:label	365430152	45.35%
Extensions: BMO, RSO	35903831	4.46%
FRs (70M=9%) and sub-FRs (26M=3%)	96526377	11.98%
Thesauri: BIBO, DC, DCT, FOAF, SKOS, QUDT, VAEM	5715250	0.71%
Ontology: RDF, RDFS, OWL	4159	0.00%
Total	**805729356**	**100.00%**
CRM inverses	149465596	18.55%

- Total 339 properties, grouped above
- Type statements take 37%: too much (see prev slides)
- Inverses (79) are convenient, but take 18% (duplicates)
- Sub-properties: max depth is 4 (e.g.: P12>P11>P14>P22).
  No estimate of the sub-property inference, sorry
- Objects take the majority: 45%
- Thesauri and ontologies are negligible: 0.7%
- FRs take only 12%, which doesn't slow OWLIM perceptibly

Comparison to Other Repositories

Repo	Objects		Expl.stat.		Ex.st/obj		Total stat.		Expans.		Nodes		Density		Reasoning
CRM	2.0	1	195	1	90	1	916	1	4.7	1	54	1	17.0	1	rdfs+tran+FR
PSNC	3.1	1.5	234	1.2	75	0.83	535	0.58	2.3	0.49	60	1.1	8.9	0.52	rdfs-subClass
EDM	20.3	9.8	998	5.1	50	0.56	3798	4.1	3.8	0.8	266	4.9	14.3	0.84	owl-horst
FF			1673	8.6			3211	3.5	1.9	0.4	456	8.4	7.0	0.41	owl-horst
LLD			6706	34			10192	11	1.5	0.3	1554	29	6.6	0.38	rdfs+tran

- **Repos:**
  - RS CRM: http://test.researchspace.org:8081
  - PSNC Polish Digital Library: http://dl.psnc.pl
  - Europeana EDM: http://europeana.ontotext.com
  - FactForge: http://www.factforge.net
  - LinkedLifeData: http://linkedlifedata.com
- **First** col is Million triples (exc. Expansion/Density), **second** col is ratio to CRM
- **Expansion**=Total statements/Explicit statements: intensity of inference
- **Nodes**=unique URIs and literals
- **Density**=Statements/Nodes: relative density of the graph

Performance: SPARQL Implementation

- Straight SPARQL 1.1 for "FR92i_created_by rkd-artist:Rembrandt":

```
select distinct ?obj {
 ?obj a rso:FC70_Thing;
  (crm:P46_is_composed_of|crm:P106_is_composed_of|crm:P148_has_component|crm:P128_carries)*/
  (crm:P31i_was_modified_by|crm:P94i_was_created_by) / crm:P9_consists_of* /
   crm:P14_carried_out_by / crm:P107i_is_current_or_former_member_of*
 rkd-artist:Rembrandt
} limit 20
```

- RS endpoint takes over 15 minutes to answer. If you add more FRs, even worse. The reflexive * really kills it
- The query can be optimized a bit by using intermediate variables instead of property paths, but the performance is still untenable

Performance: our Implementation

- Objects by Rembrandt: sub-second response time:
  select distinct ?obj {?obj rso:FR92i_created_by rkd-artist:Rembrandt} limit 500
- Find terms "drawing" and "mammal":
  select * {?s rdfs:label "drawing"} → thes:x6544
  select * {?s rdfs:label "mammal"} → thes:x12965
- **Drawings** by Rembrandt about **mammals**: still sub-second response time, and the query is simple:

```
select distinct ?obj {
 ?obj rso:FR92i_created_by rkd-artist:Rembrandt;
    rso:FR2_has_type thes:x6544, thes:x12965} limit 500
```

- RS search takes 4.5s (significantly longer than the query alone) because after obtaining up to 500 objects, it executes several more queries to fetch their display fields, facets, and images
- Facets are loaded into the browser using Exhibit, so subsequent facet restrictions are immediate

Thanks for listening!

- Questions? vladimir.alexiev@ontotext.com