EXPLOITING SEMANTICS FOR BIG DATA INTEGRATION

WHAT IS THIS ARTICLE ABOUT?

- Exploiting semantics to solve the problem of Big Data Variety
- An approach to integrate data from multiple types of sources
  - Spreadsheet
  - Relational databases
  - Web services
  - We address these variety using **Karma**

KARMA

Main benefits

- Allows import from a wide variety of sources
- Clean and normalize data
- Quickly build a model or sematic description of each source
- Integrate the data across sources using that model

Being used in..

- Biological data
- Phono data
- Geospatial data
- Cultural heritage data
- Environmental data

STRUCTURE

1. Importing
2. Cleaning
3. Modeling
4. Integrating data

1. Problems Karma are still trying to fix

Challenges

&

Solutions

IMPORTING

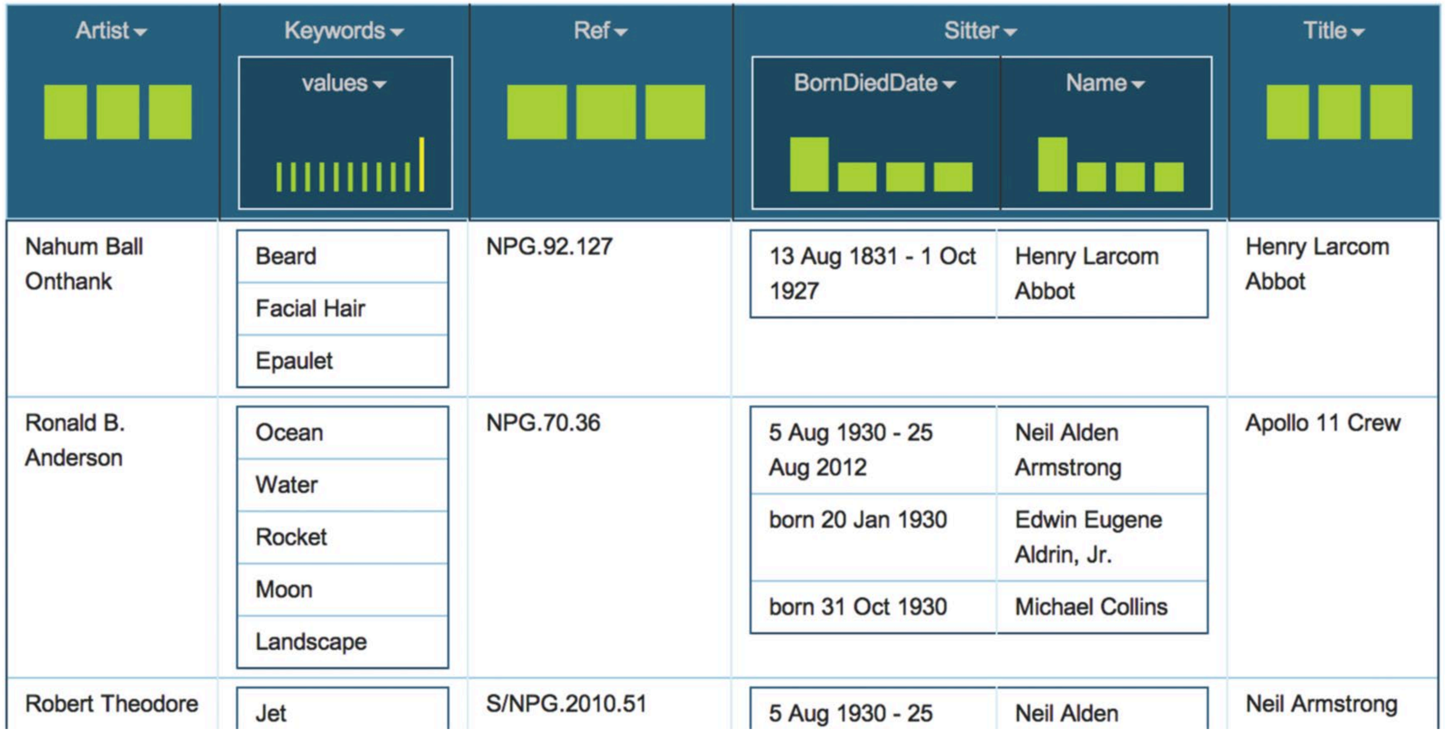
Challenges

- Importin different data formats into a common representation
- When the sources are large it is not possible to read an entire source into main memory

1. IMPORTING

Solution

- Converting all data formats into a nested relational data model
- Data is represented in tables where cells can contain scalar values
- Karma imports XML documents similarly



Artist	Keywords	Ref	Sitter		Title
	values		BornDiedDate	Name	
Nahum Ball Onthank	Beard Facial Hair Epaulet	NPG.92.127	13 Aug 1831 - 1 Oct 1927	Henry Larcom Abbot	Henry Larcom Abbot
Ronald B. Anderson	Ocean Water Rocket Moon Landscape	NPG.70.36	5 Aug 1930 - 25 Aug 2012 born 20 Jan 1930 born 31 Oct 1930	Neil Alden Armstrong Edwin Eugene Aldrin, Jr. Michael Collins	Apollo 11 Crew
Robert Theodore	Jet	S/NPG.2010.51	5 Aug 1930 - 25	Neil Alden	Neil Armstrong

2. CLEANING

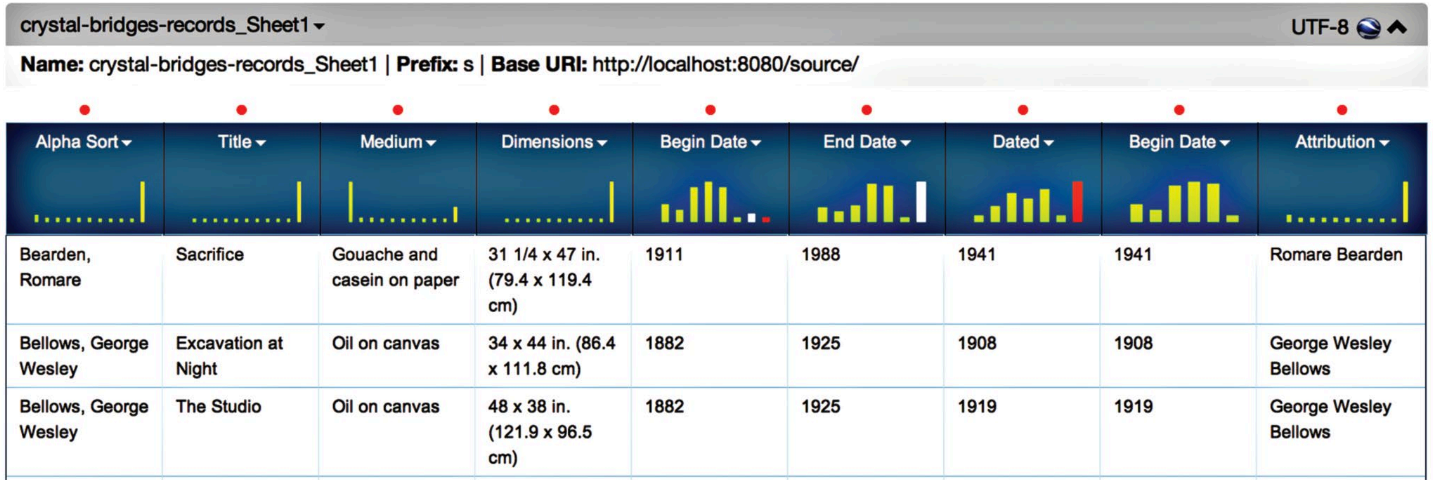
Challenges

- Noisy data, missing values, and inconsistencies that need to be identified and fixed
- The data in different sources is often represented in different and incompatible ways

2. CLEANING

Solution

- Karma helps find the inconsistent data by performing an analysis of the data distribution in each Colum
  - The white bar shows the null values
  - The red bar shows the frequency of outliers



crystal-bridges-records_Sheet1 ▾ UTF-8

Name: crystal-bridges-records_Sheet1 | **Prefix:** s | **Base URI:** http://localhost:8080/source/

Alpha Sort ▾	Title ▾	Medium ▾	Dimensions ▾	Begin Date ▾	End Date ▾	Dated ▾	Begin Date ▾	Attribution ▾
Bearden, Romare	Sacrifice	Gouache and casein on paper	31 1/4 x 47 in. (79.4 x 119.4 cm)	1911	1988	1941	1941	Romare Bearden
Bellows, George Wesley	Excavation at Night	Oil on canvas	34 x 44 in. (86.4 x 111.8 cm)	1882	1925	1908	1908	George Wesley Bellows
Bellows, George Wesley	The Studio	Oil on canvas	48 x 38 in. (121.9 x 96.5 cm)	1882	1925	1919	1919	George Wesley Bellows

3. MODELING

Challenges

- One of the main challenges of integrating diverse data sets is to harmonize their representation
- Nomenclature differences: Data sets from different providers often use different names to refer to attributes that have the same meaning.
- Format and structure differences: Different data sets come in different formats.

3. MODELING

Solution

- In Karma, they address these differences by modeling all the data sets with respect to a common ontology
- This involves two steps
  - Assignment of sematic types to data columns
  - Specification of the relationships between the semantic types
- Learning from previously defined models
- Learning coherent substructures
- **This model greatly reduces the effort needed to create new models**

4. INTEGRATING DATA

- **Challenges**

Involves 2 steps

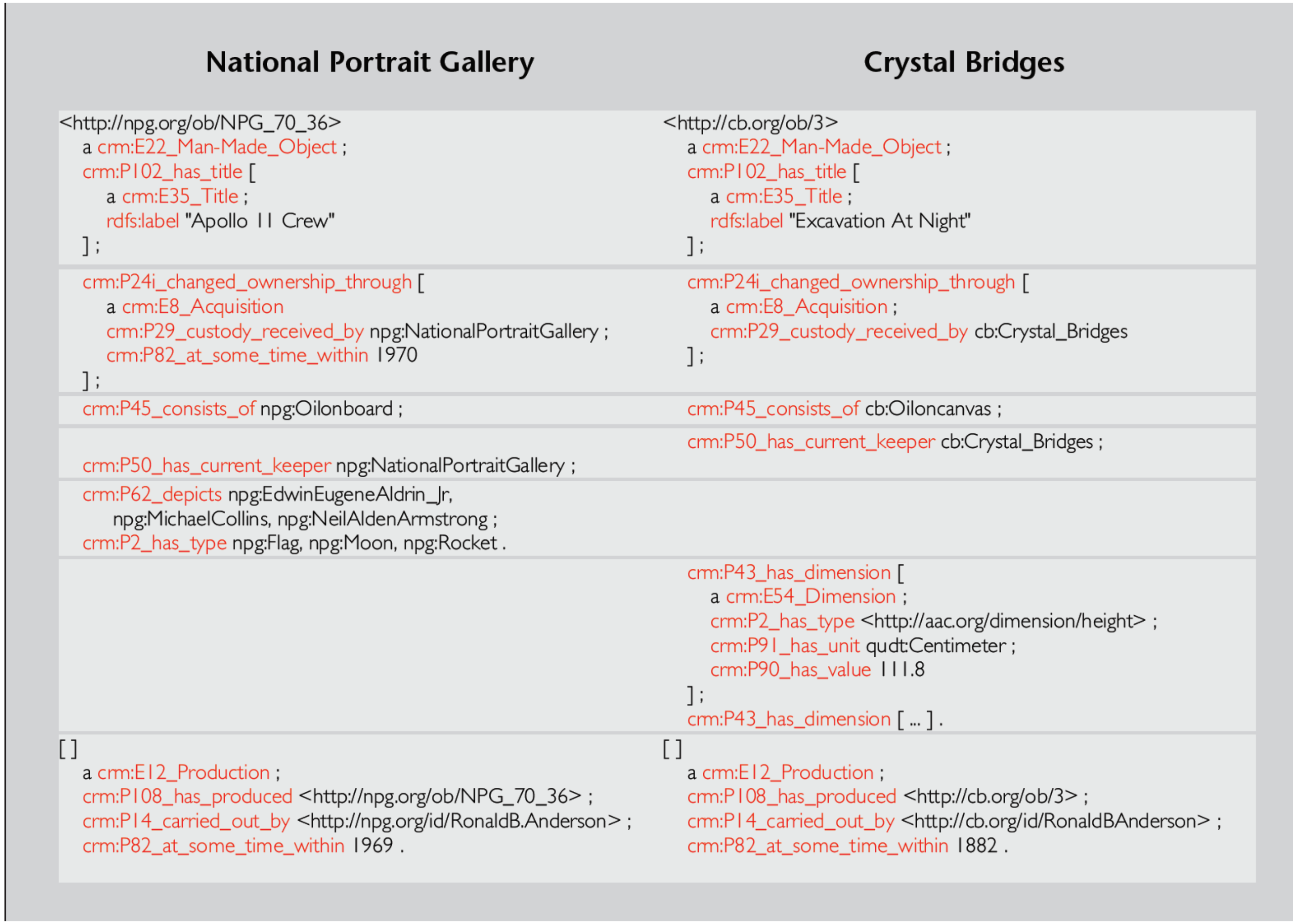
1. At the schema level, it involves homogenizing differences in t**he schemas** and nomenclature used to represent the data.
2. The second integration at the data level involves identifying records in different data sets that refer to the same real-world entity

- **Karma focuses on the schema level integration problem**

4. INTEGRATING DATA

Karmas solution

- Using a common domain ontology
  - Once the user models them using the CRM ontology, Karma can convert the data into RDF using a common set of terms
- Museum example:
  - Can be easily queried using SPARQL
  - Karma can also convert the data to CSV
  - **The advantage is that the converted data up to date from the database**



National Portrait Gallery

```
<http://npg.org/ob/NPG_70_36>
  a crm:E22_Man-Made_Object ;
  crm:P102_has_title [
    a crm:E35_Title ;
    rdfs:label "Apollo 11 Crew"
  ] ;
  crm:P24i_changed_ownership_through [
    a crm:E8_Acquisition
    crm:P29_custody_received_by npg:NationalPortraitGallery ;
    crm:P82_at_some_time_within 1970
  ] ;
  crm:P45_consists_of npg:Oilonboard ;
  crm:P50_has_current_keeper npg:NationalPortraitGallery ;
  crm:P62_depicts npg:EdwinEugeneAldrin_Jr,
     npg:MichaelCollins, npg:NeilAldenArmstrong ;
  crm:P2_has_type npg:Flag, npg:Moon, npg:Rocket .
[ ]
  a crm:E12_Production ;
  crm:P108_has_produced <http://npg.org/ob/NPG_70_36> ;
  crm:P14_carried_out_by <http://npg.org/id/RonaldB.Anderson> ;
  crm:P82_at_some_time_within 1969 .
```

Crystal Bridges

```
<http://cb.org/ob/3>
  a crm:E22_Man-Made_Object ;
  crm:P102_has_title [
    a crm:E35_Title ;
    rdfs:label "Excavation At Night"
  ] ;
  crm:P24i_changed_ownership_through [
    a crm:E8_Acquisition ;
    crm:P29_custody_received_by cb:Crystal_Bridges
  ] ;
  crm:P45_consists_of cb:Oiloncanvas ;
  crm:P50_has_current_keeper cb:Crystal_Bridges ;
  crm:P43_has_dimension [
    a crm:E54_Dimension ;
    crm:P2_has_type <http://aac.org/dimension/height> ;
    crm:P91_has_unit qudt:Centimeter ;
    crm:P90_has_value 111.8
  ] ;
  crm:P43_has_dimension [ ... ] .
[ ]
  a crm:E12_Production ;
  crm:P108_has_produced <http://cb.org/ob/3> ;
  crm:P14_carried_out_by <http://cb.org/id/RonaldBAnderson> ;
  crm:P82_at_some_time_within 1882 .
```

KARMA IS NOT PERFECT…

- Karma focuses on integrating sources at the schema level, but there is an equally important problem of linking the data at he record level.
- The article focuses on the issue of variety and did not address the issues of volume and velocity, which are the other key dimensions of big data.