# How Portable Are the Metadata Standards for Scientific Data?

Jian Qin Kai Li
School of Information Studies
Syracuse University
Syracuse, NY, USA

# Why study metadata portability?

**Complex, very large metadata standards**

- are "…unwieldy to apply…"
- are "difficult to understand and enact in its entirety…"
- require customization to tailor to specific needs
- costly in time and personnel





# Each standard has its own schema and tools…


EPA Metadata Editor

Metavist2


Morpho


AVM Tagging Tool


Protocol Registration System

…that lead to duplicated efforts and interoperability problems

# Metadata for scientific data is at the juncture of -

## Data-Driven Science

## Technical Standards Infrastructure

RDFa

RDF OWL ORCID

XML URI DOI

# A few big questions

- What action should and can we take at this juncture as a community of metadata practices?
- How much do we know about metadata standards for scientific data?
- How can we transform the current metadata standards into an infrastructure-driven service?

# An infrastructure perspective for metadata

- Portable
- Customizable
- Extendable
- Reusable
- Easy to use

# An attempt to define “metadata infrastructure”

**Semantically:**

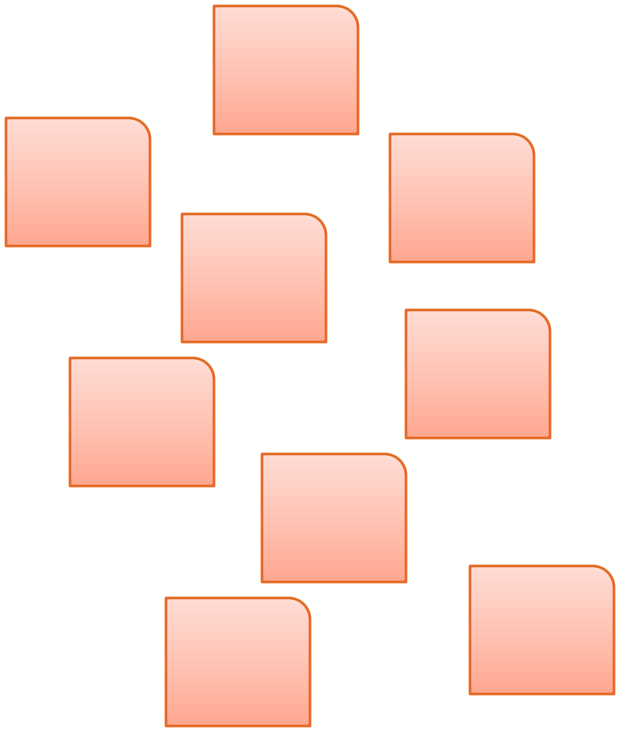
- metadata elements, vocabularies, entities, and other metadata artifacts as the underlying foundation to build tools, software and applications

**Technically:**

- “a data model for describing the resources, aspects of metadata encoding and storage formats, metadata for web services, metadata tools, usage, modification, transformation, interoperability, and metadata crosswalk ” (CLARIN)

# Portability is the key

**Building blocks of metadata**



**Metadata generation output**

- Metadata for data citation
- Metadata for data discovery
- Metadata for data archiving
- Metadata for data quality
- Metadata for data provenance
- Metadata for data management

# How portable are metadata standards for scientific data?

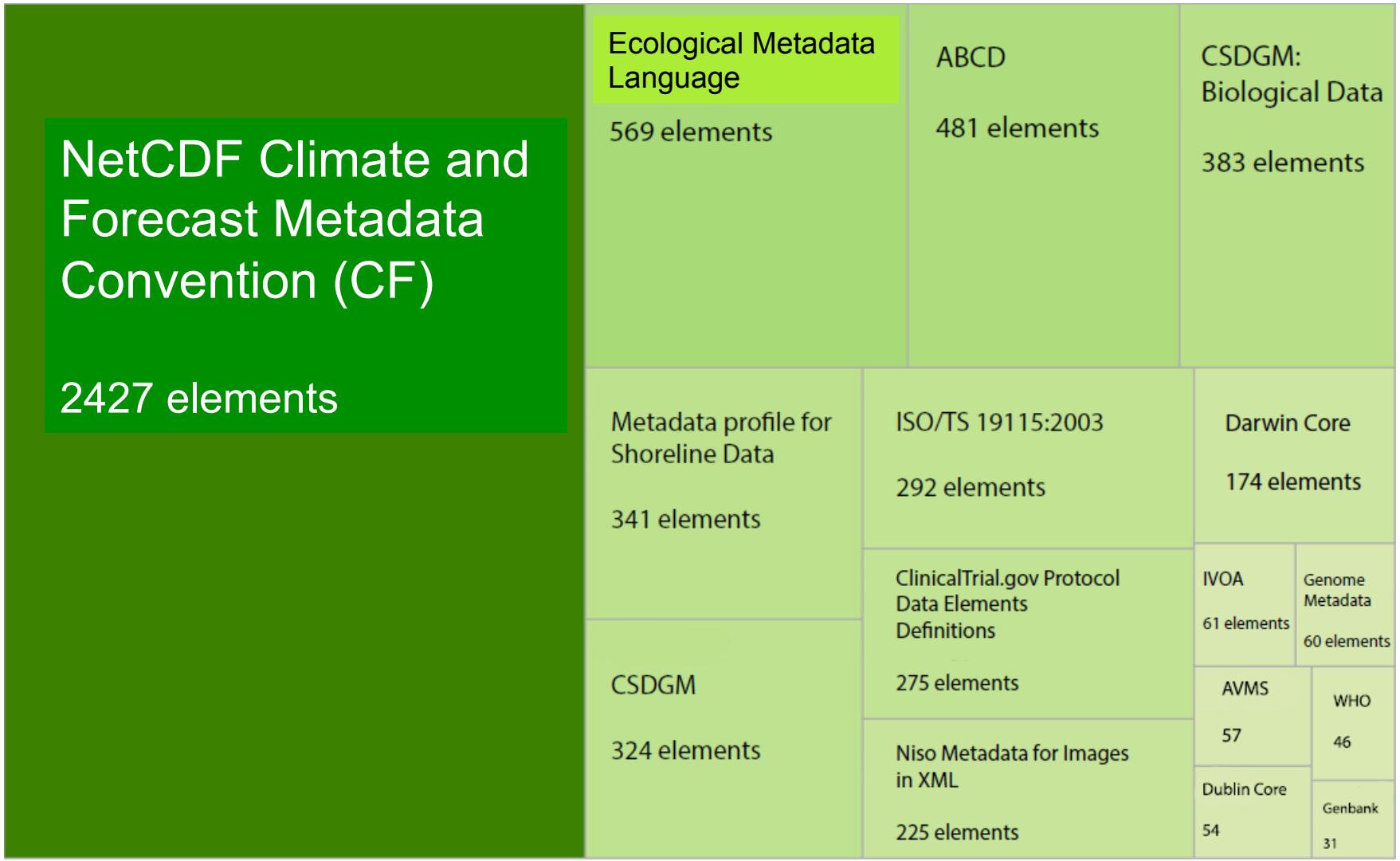
**Two measures of metadata portability:**

- Co-occurrence of semantic elements: the times of semantically identical elements used in multiple standards
- Degree of modularity: the degree of independence and self-descriptiveness of a sub-structure of concept/entity in metadata standards

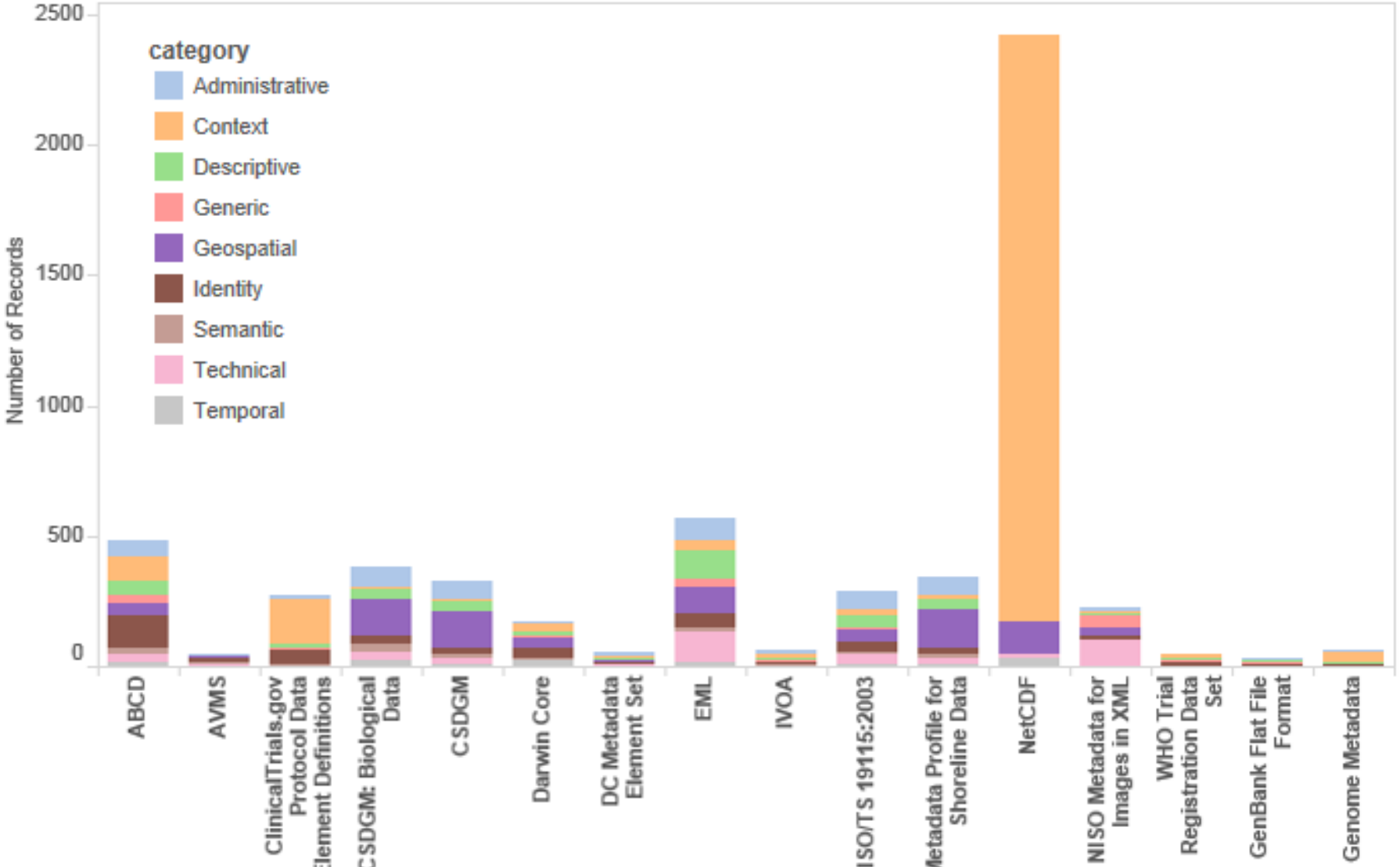
# Data

| | |
|---|---|
| Element Collection | • 5,800 elements from 16 scientific metadata standards |
| Element De-duplication | • 4,434 unique elements in terms of semantic |
| Categorization | • 9 categories based on functionalities of the elements |

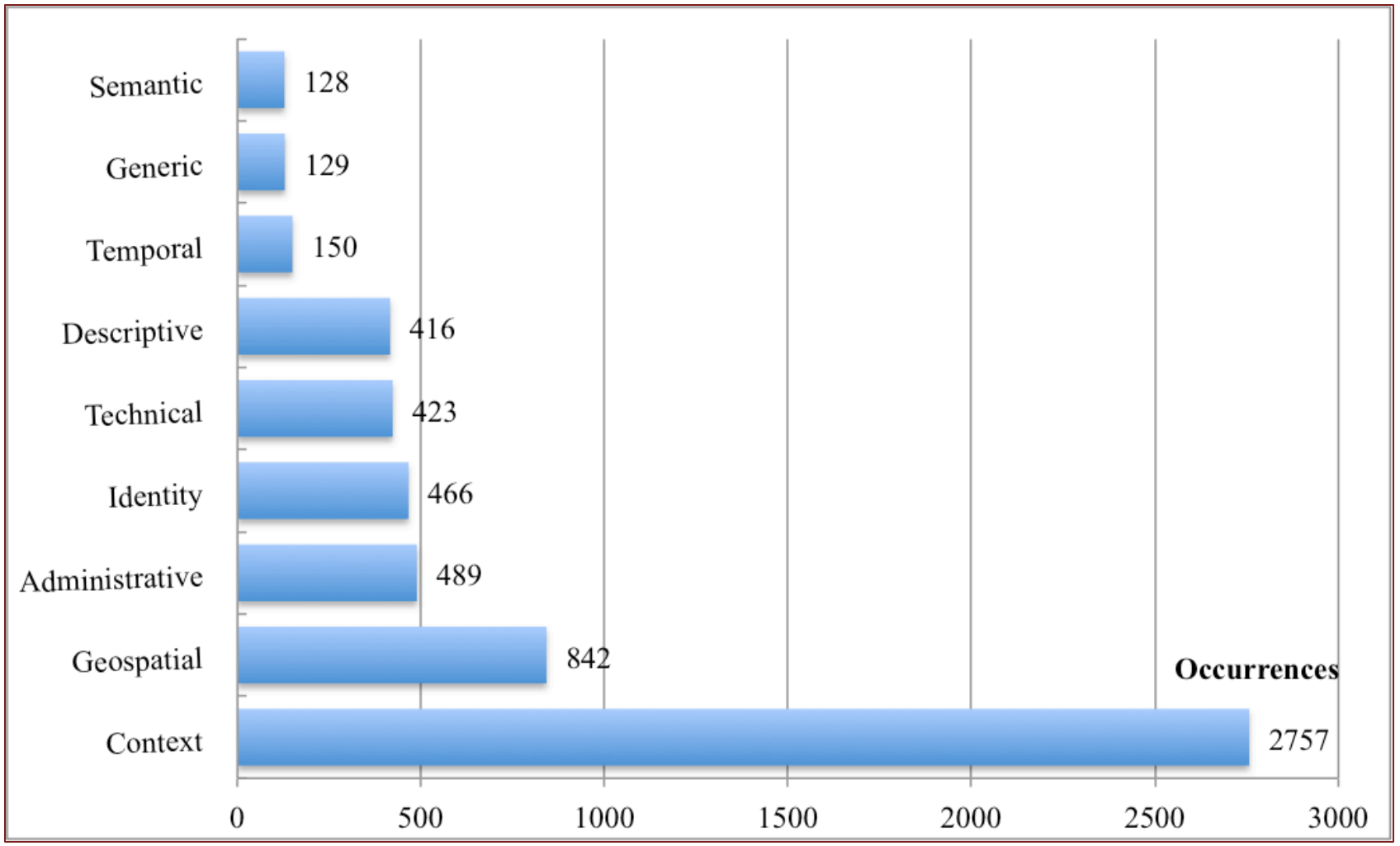
# Element distribution by standard



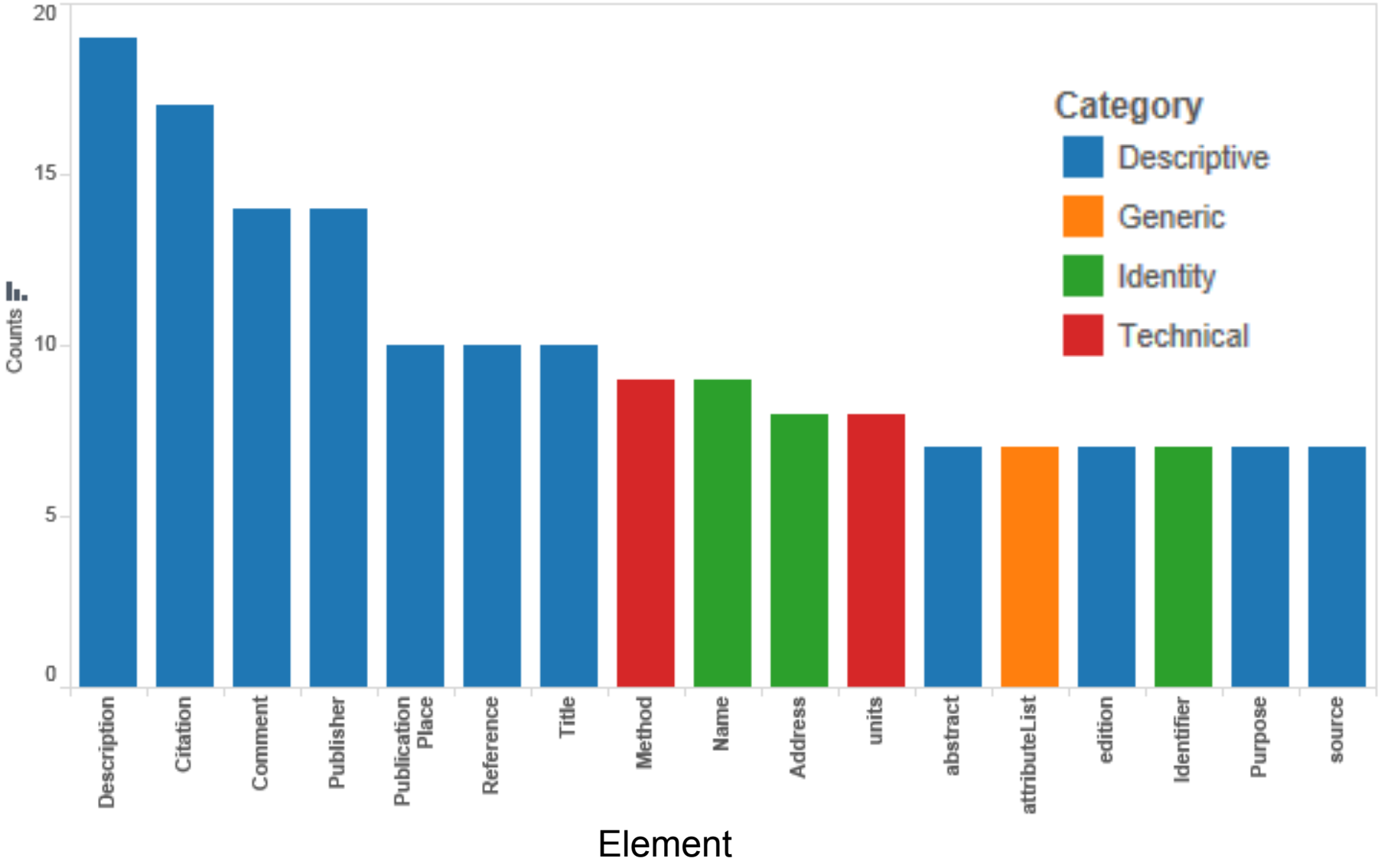
# Element distribution by standard and catagory



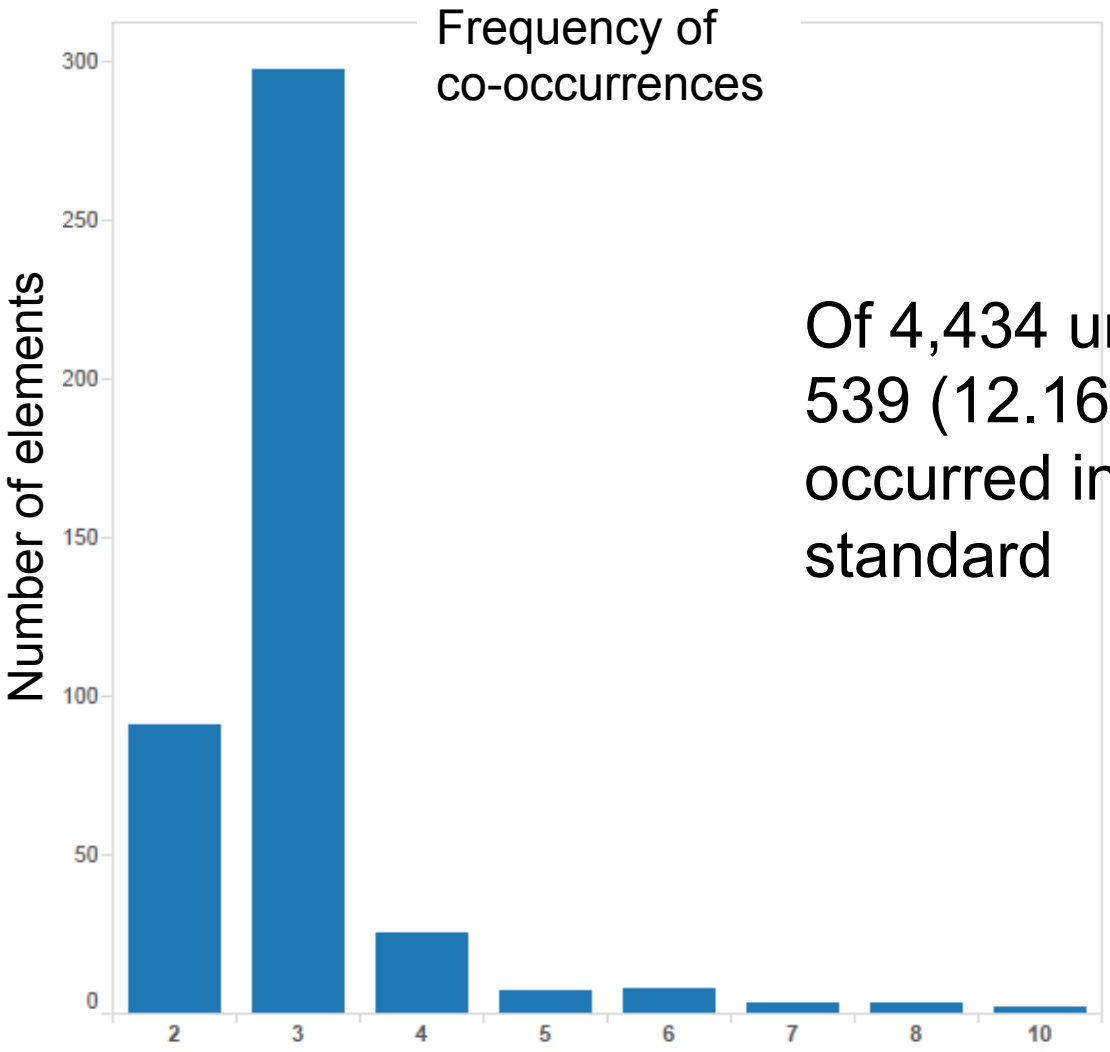
# Frequency of occurrences by category



# Top occurring elements

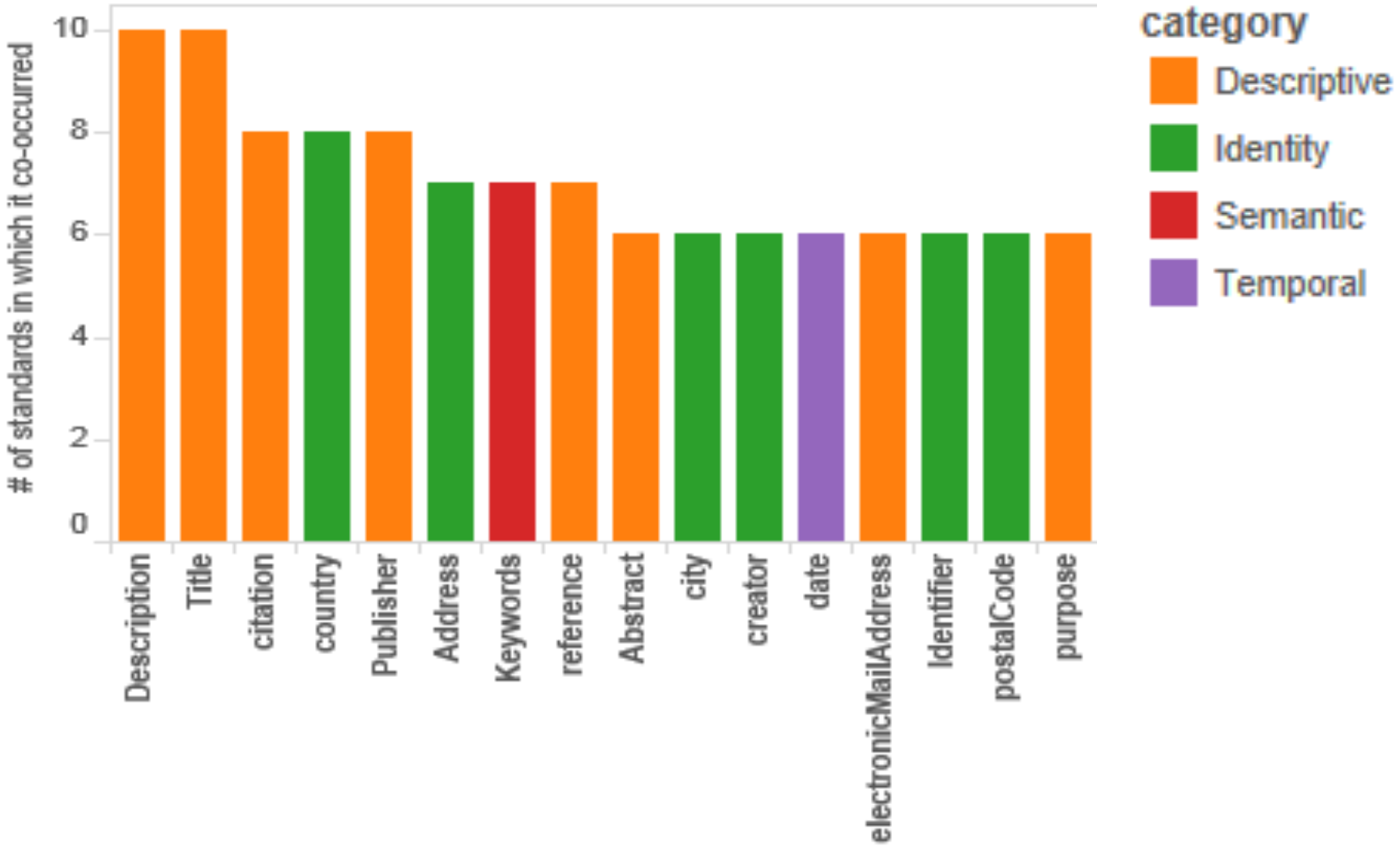


# Element co-occurrences across metadata standards




Of 4,434 unique elements, 539 (12.16%) elements occurred in more than one standard

# Elements that most frequently co-occurred



# Modularity

**Two levels of modularity:**

- Level 1: having multiple XML schema files for the whole standard;
- Level 2: having separate schemas for entities such as person/organization, dataset, study, instrument, and subject

**Of the 6 standards with schema files, all of them belong to Level 1 modularity.**

# Discussion

**Portable metadata standards**

- Possible?
- Feasible?
- Advantages over the one-covers-all approach?

**A metadata infrastructure for scientific data**

- Bridge the gap between existing semantic and entity resources and metadata generation
- Much to be researched…

# Further research

**More questions than answers from this study:**

- What should a metadata infrastructure constitute?
- How can the gaps be filled or narrowed between the infrastructure resources and metadata applications?
- Is it possible or is there a need to streamline the metadata scheme design practice toward a metadata infrastructure?
- …and the list can go on

# Questions and comments?

# Name length of elements by category