INTEROPERABILITY

What it means and what we need it to do

Peter Winstanley
03 September 2018

Lack of interoperability results in **cost** or **danger**

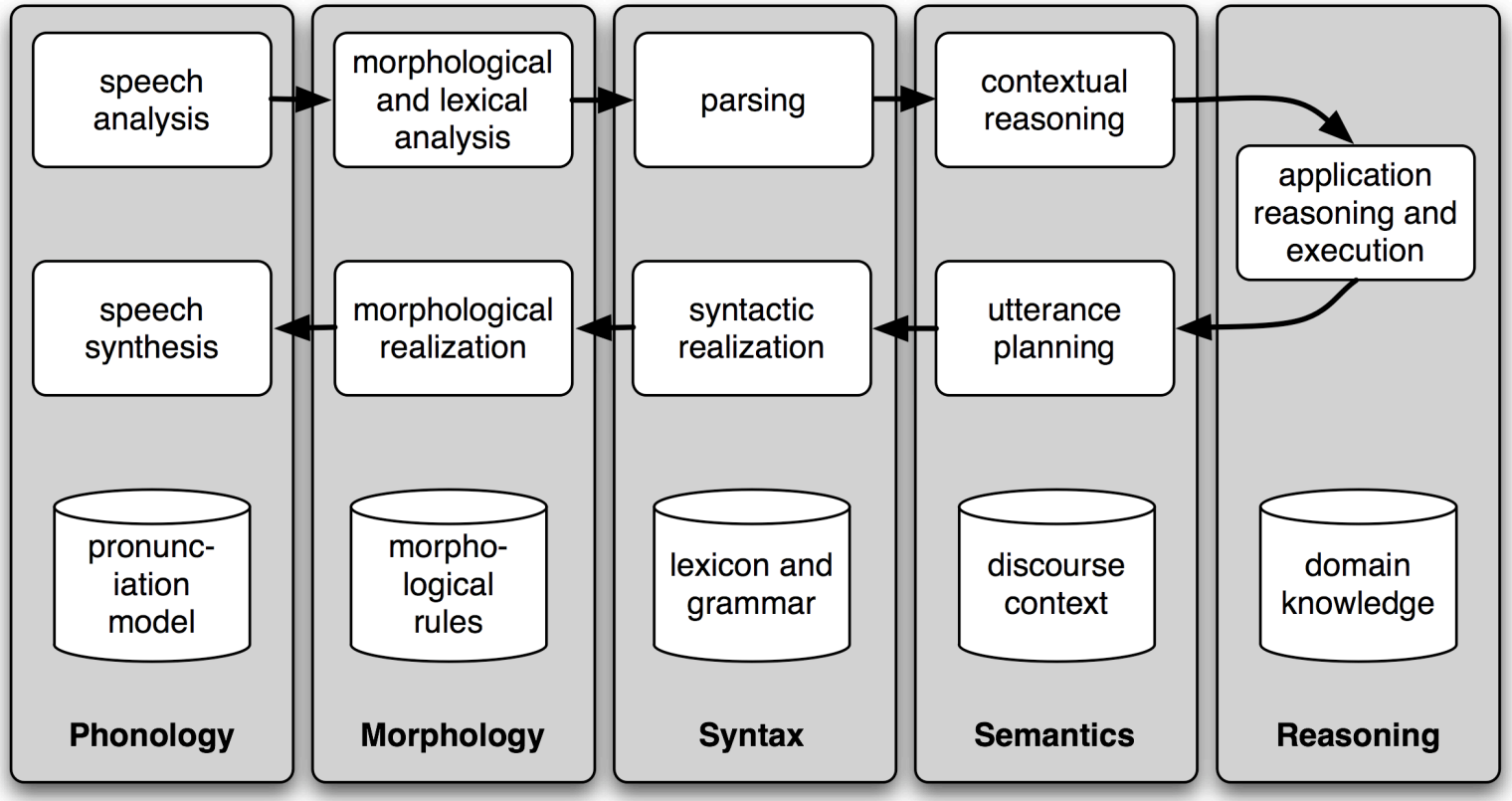
A *shared language* is practical and important

Main components of spoken language

- A **phoneme** is the smallest unit of sound that may cause a change of meaning within a language but that doesn't have meaning by itself.
- A **morpheme** is the smallest unit of a word that provides a specific meaning to a string of letters (which is called a phoneme).
- A **lexeme** is the set of all the inflected forms of a single word. A basic abstract unit of meaning.
- **Syntax** is the set of rules by which full sentences are constructed
- **Context** is how everything within language works together to convey a particular meaning [i.e. semantics]

Main components of **written** language

- A **grapheme** is the smallest unit of writing that represents a sound (e.g. individual letters, diphthongs, or combinations such as “igh”)
- A **morpheme** is the smallest unit of a word that provides a specific meaning to a string of letters (phoneme or grapheme).
- A **lexeme** is the set of all the inflected forms of a single word. A basic abstract unit of meaning.
- **Syntax** is the set of rules by which full sentences are constructed
- **Context** is how everything within language works together to convey a particular meaning [i.e. semantics]



Our communication pathway

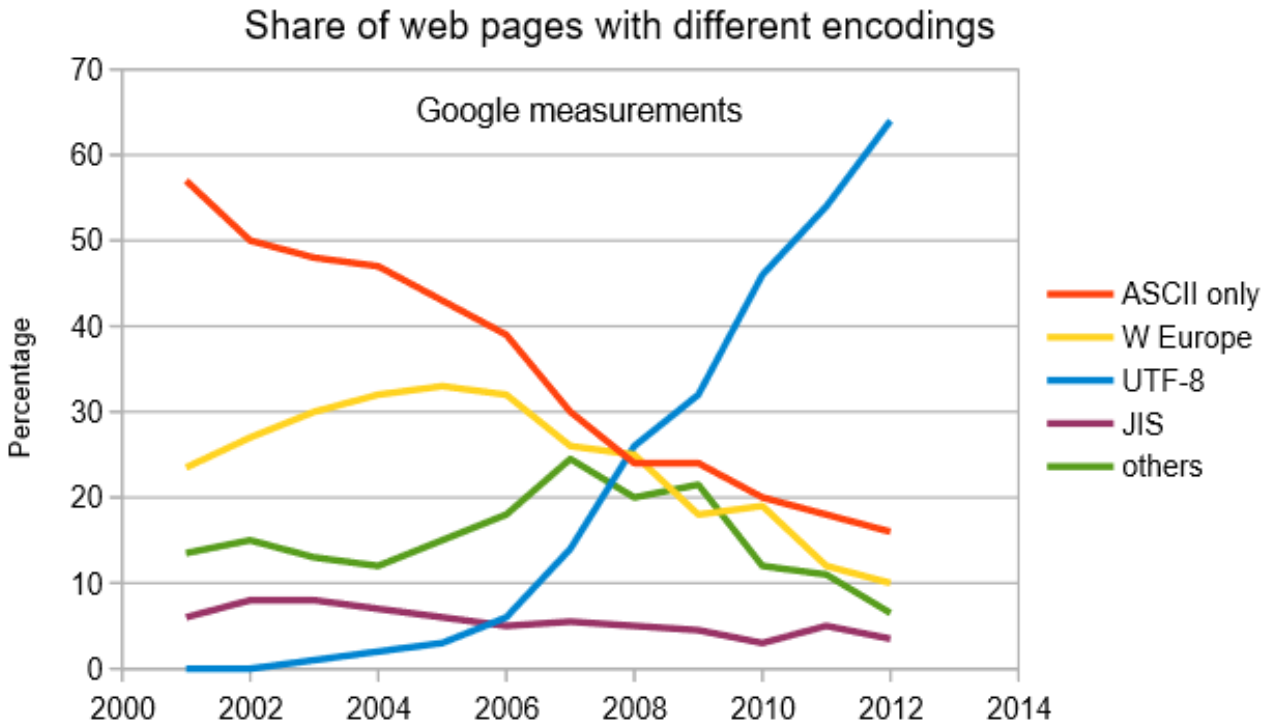
How does **morphology** apply to computers?

Human	Computer
Character set [e.g. “abcd” vs ”αβγδ”]	Character set [e.g. “asci” vs “utf-8”]

In natural written language we have different scripts to represent the sounds we make in speech. Computers use ‘encoding’ and ‘character sets’ for the same purpose. Using the same encoding speeds up interoperability. It makes things much easier.

How does **morphology** apply to computers? (cont.)

UTF-8 has been the dominant character encoding for the World Wide Web since 2009, and as of August 2018 accounts for 92.0% of all web pages and 95.5% of the top 1,000 highest ranked web pages. W3C recommends UTF-8 as the default encoding in XML and HTML.



Recommendation 1: Use UTF-8 for character encoding

How does **parsing** apply to computers?

In natural written language we have pauses and inflection in the sounds we make in speech to separate words and sentences. In written text we use punctuation and paragraphs. Computers use markup for the same purpose [1]. Using the same markup speeds up interoperability. It makes things much easier.

Recommendation 2: Use a flexible, known machine-readable format specification such as XML or JSON. [http://www.w3.org/TR/dwbp/#dataFormats]

The eXtensible Mark Language (XML) is a set of syntactic rules that allow users to develop their own annotations, and more precisely, their own markup languages. The ability to easily create, read and modify one's own document structures using a standard template facilitates syntactic interoperability. In the case of XML, this is achieved through a hierarchical structure composed of elements and attributes. Alternatively, JSON adopts a less cluttered format based on attribute-key-value pairs to

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4434901/

What does **semantic interoperability** mean for computers?

The adoption of a generalized markup language such as XML is not sufficient to create interoperability.

Both sides of a communication need to be able to interpret the information exchanged. To achieve semantic interoperability, controlled vocabularies, and standard taxonomies are required.

A more generic solution is achieved through the adoption of ontology languages. Such Semantic Web techniques allow the encoding of knowledge about specific domains by augmenting existing documents with attributes that denote meaning.

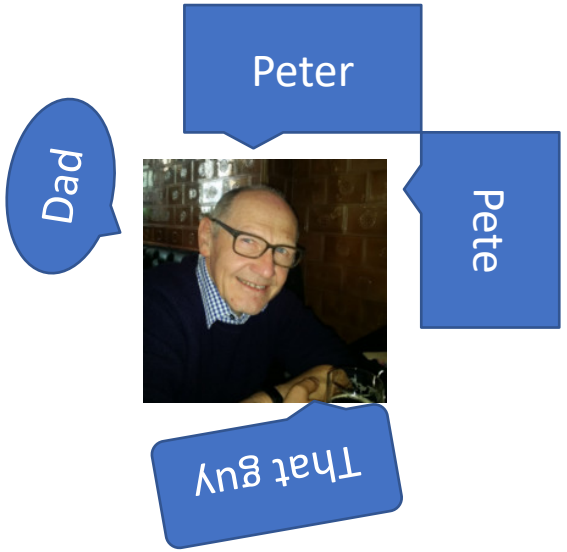
This allows information to be exchanged meaningfully and accurately, even when terms are expressed in different languages, or when two or more terms refer to the same concept but are not easily recognized as synonyms.

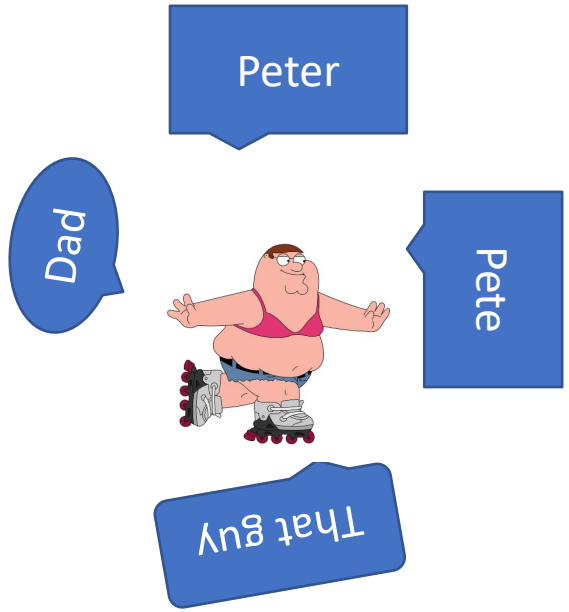
What does **semantic interoperability** mean for computers?

A key point is to make sure the dataset, or its documentation, provides enough (human- and machine-readable) context so that data consumers can retrieve and exploit the standardized meaning of the values. In the context of the Web, using unambiguous, Web-based identifiers (URIs) for standardized vocabulary resources is an efficient way to do this, noting that the same URI may have multilingual labels attached for greater cross-border interoperability. The European Union's multilingual thesaurus, Eurovoc, provides a prime example

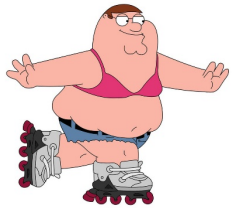
Recommendation 3: Use terms from shared vocabularies, preferably standardized ones, to encode data and metadata. [http://www.w3.org/TR/dwbp/#dataVocabularies]

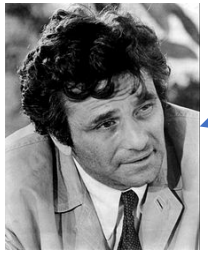
Is it good practice to use labels, such as names, as identifiers when working with digital information?





Labels are not enough.
They are **not** good identifiers because they can be reused ambiguously.





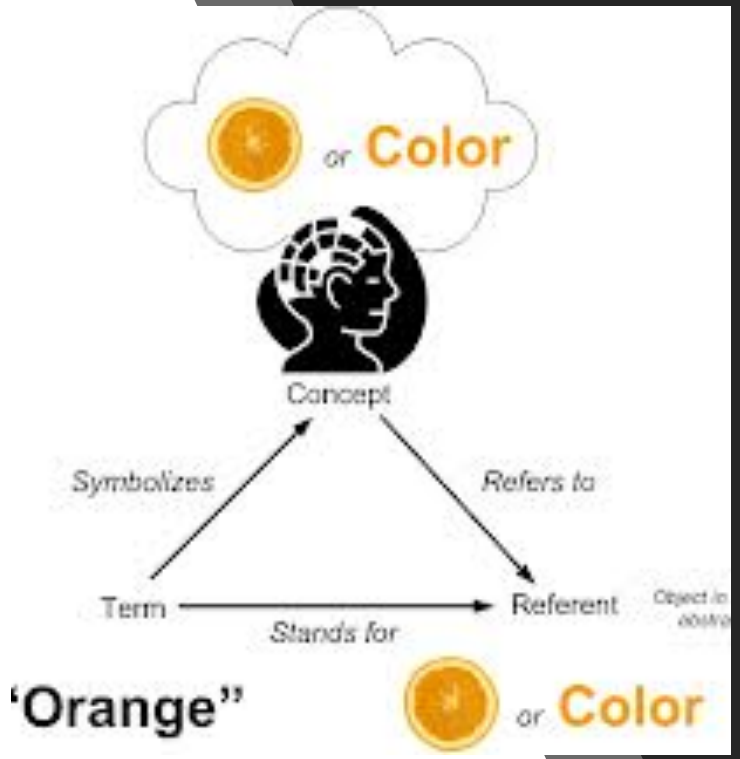
Dad

That guy

Peter

Pete

How can computers get a ‘triangle of reference’?



- The **triangle of reference** is a model of how linguistic symbols are related to the objects they represent. The triangle was published in The Meaning of Meaning by Ogden and Richards.
- Computers can share **systems of persistent identifiers** to ensure that facts or statements can be assessed as referring to the same ‘thing’. Identifiers take many forms and are used extensively in every information system. Adopting a common identification system enables basic data identification and comparison processes by any stakeholder in a reliable way. They are an essential pre-condition for proper data management and reuse.

Recommendation 4: Use persistent URIs as identifiers [http://www.w3.org/TR/dwbp/#DataIdentifiers]

- Options include HTTP(S) URIs and Digital Object Identifiers (DOI)

Next Steps: #1 - Identifiers

- Policy - http URLs as persistent (immutable) identifiers
- ‘minting’ definitive identifiers for things ⇨
  - vocabulary of things
  - vocabulary of terms

- Use VocBench and Skosmos to create and publish terminology as SKOS Concept Schemes
- see http://51.15.194.251/Skosmos/en/

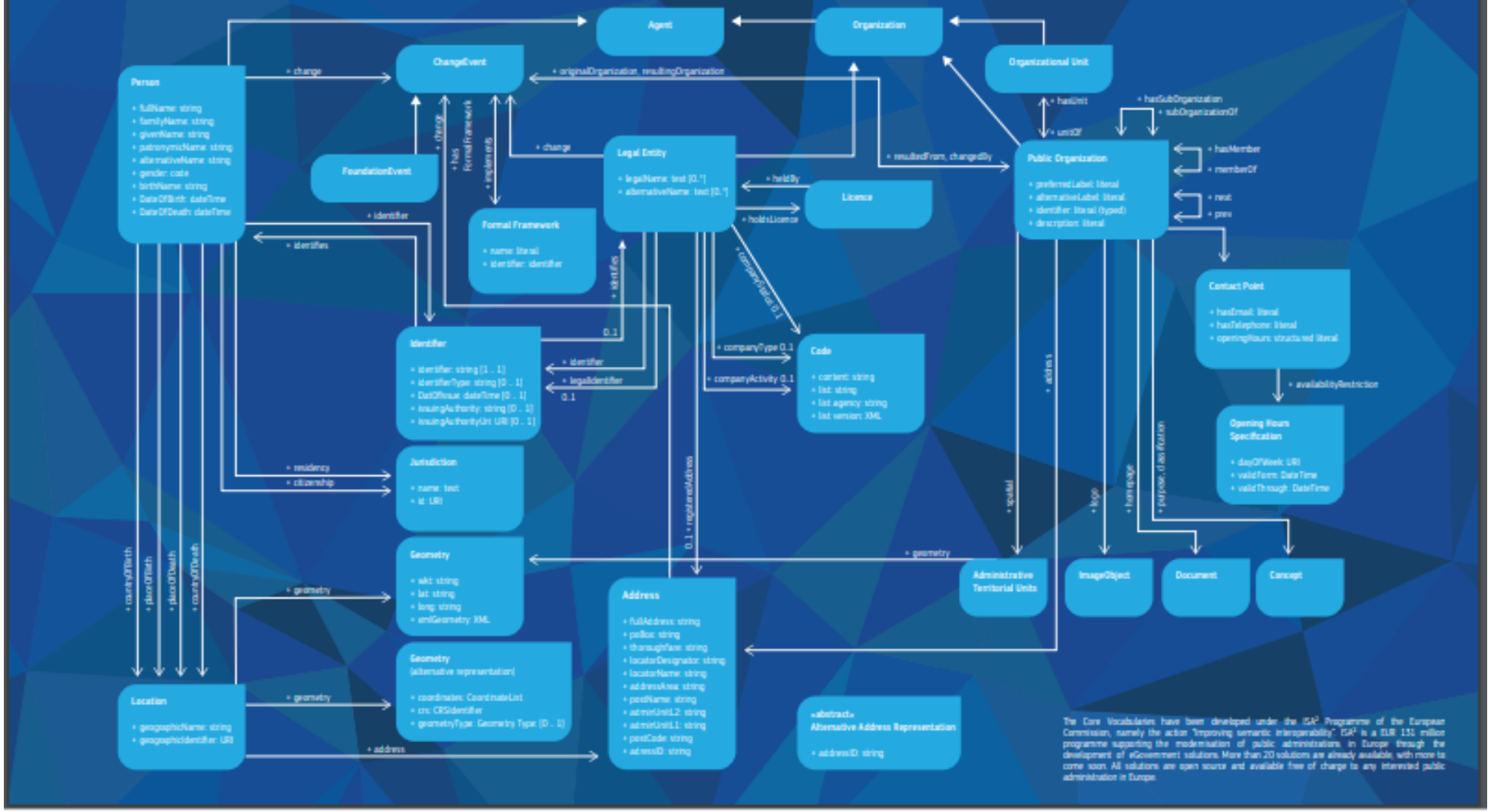
Next Steps: #2 - Terms

Next Steps: #3 - Models

- Use “Core Vocabularies” and Ontologies to share metamodels
- See https://pwin.github.io/CPOV-AP-SCOT/index-en.html
- See https://ec.europa.eu/isa2/sites/isa/files/corevocabularies-poster.pdf

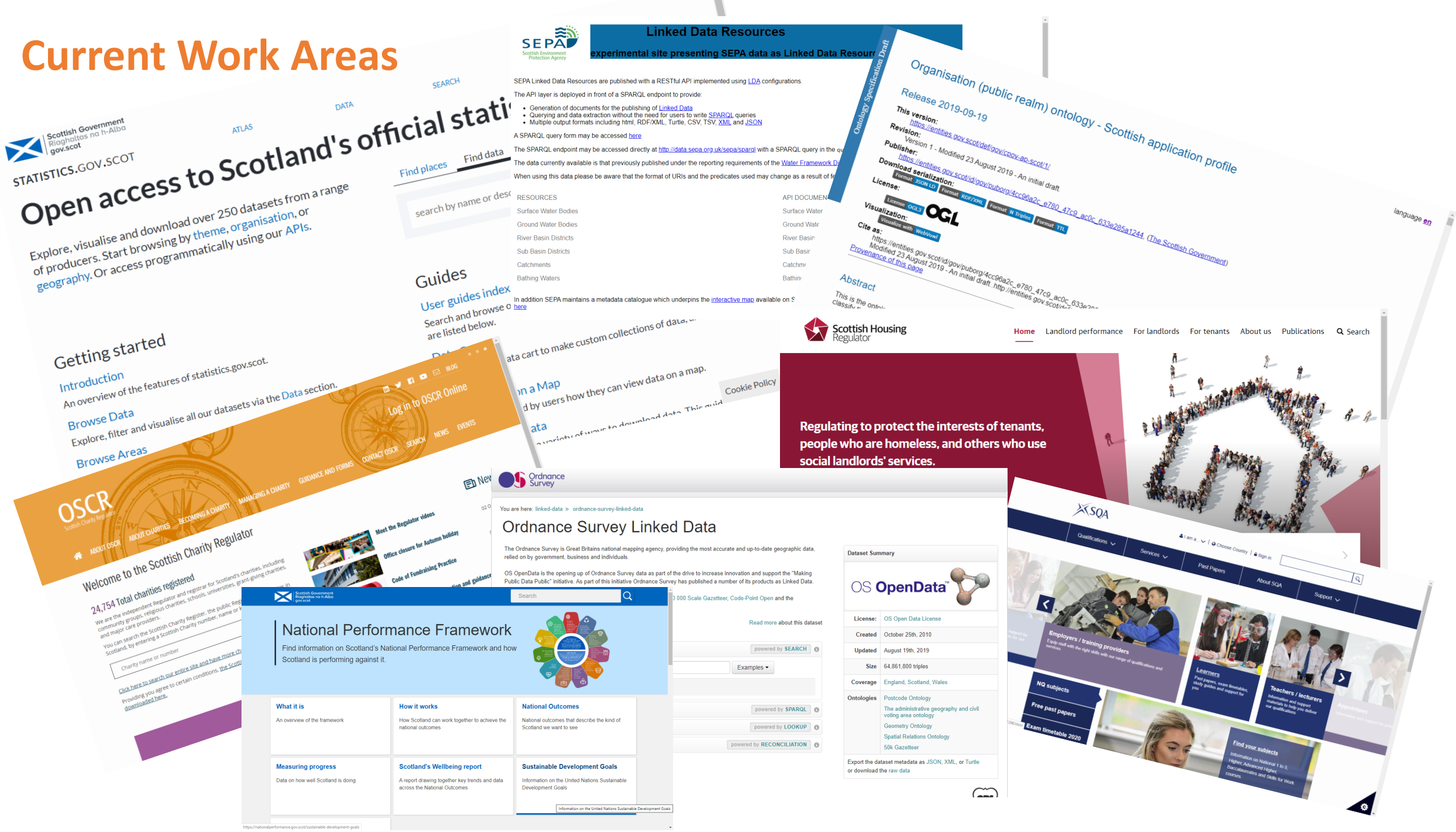
European Commission

ISA² Core Vocabularies



The Core Vocabularies have been developed under the ISA² Programme of the European Commission, namely the action "Improving semantic interoperability". ISA² is a EUR 131 million programme supporting the modernisation of public administrations in Europe through the development of eGovernment solutions. More than 20 solutions are already available, with more to come soon. All solutions are open source and available free of charge to any interested public administration in Europe.

Current Work Areas



What can we do with this?