

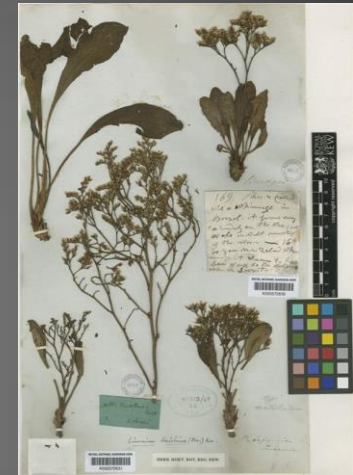
Using Complementarity to Improve Plant Specimen Digitization



Elspeth Haston
Royal Botanic Gardens-Edinburgh



Rusty Russell
Smithsonian Institution

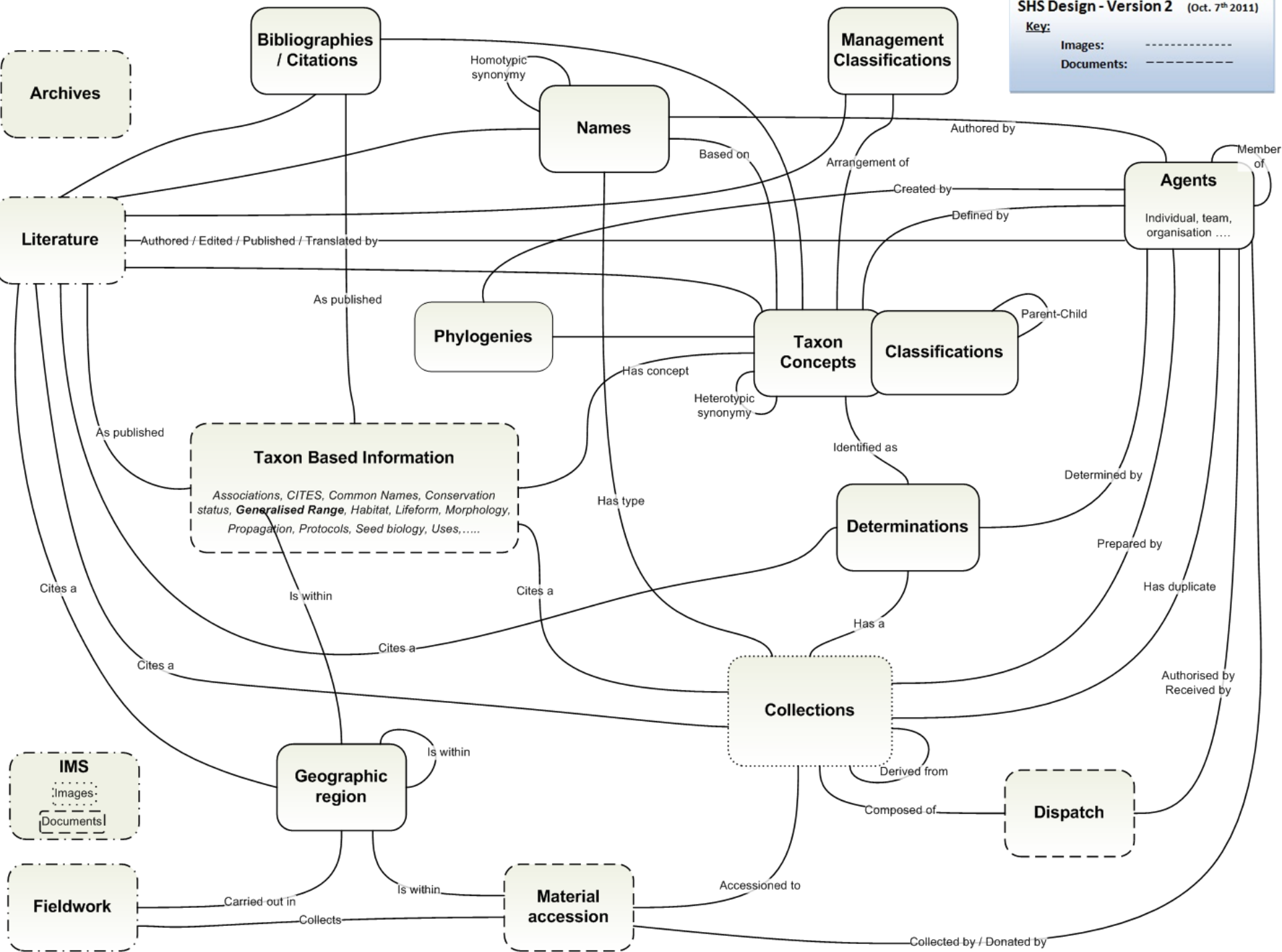


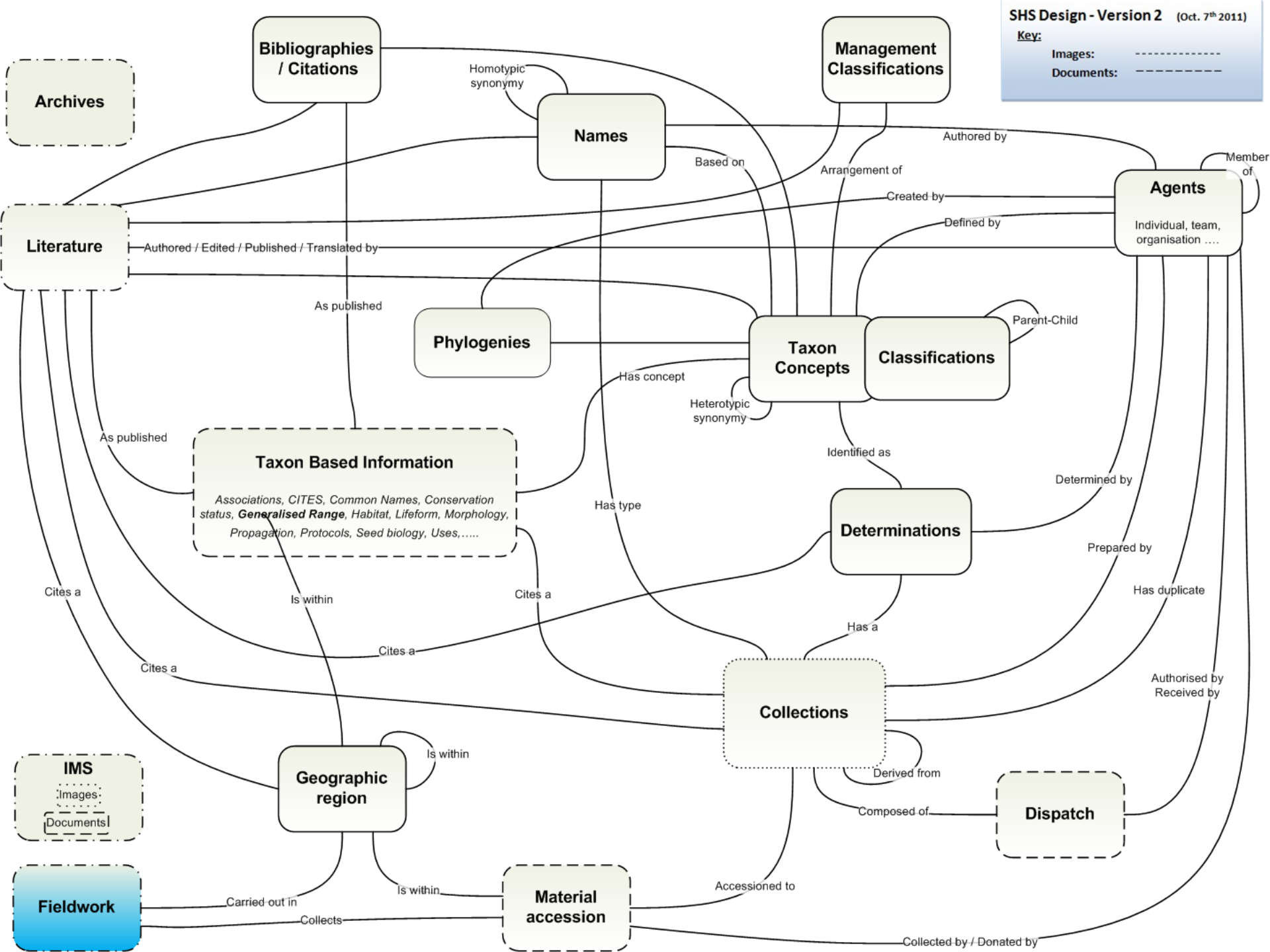
Nicola Nicolson
Royal Botanic Gardens-Kew

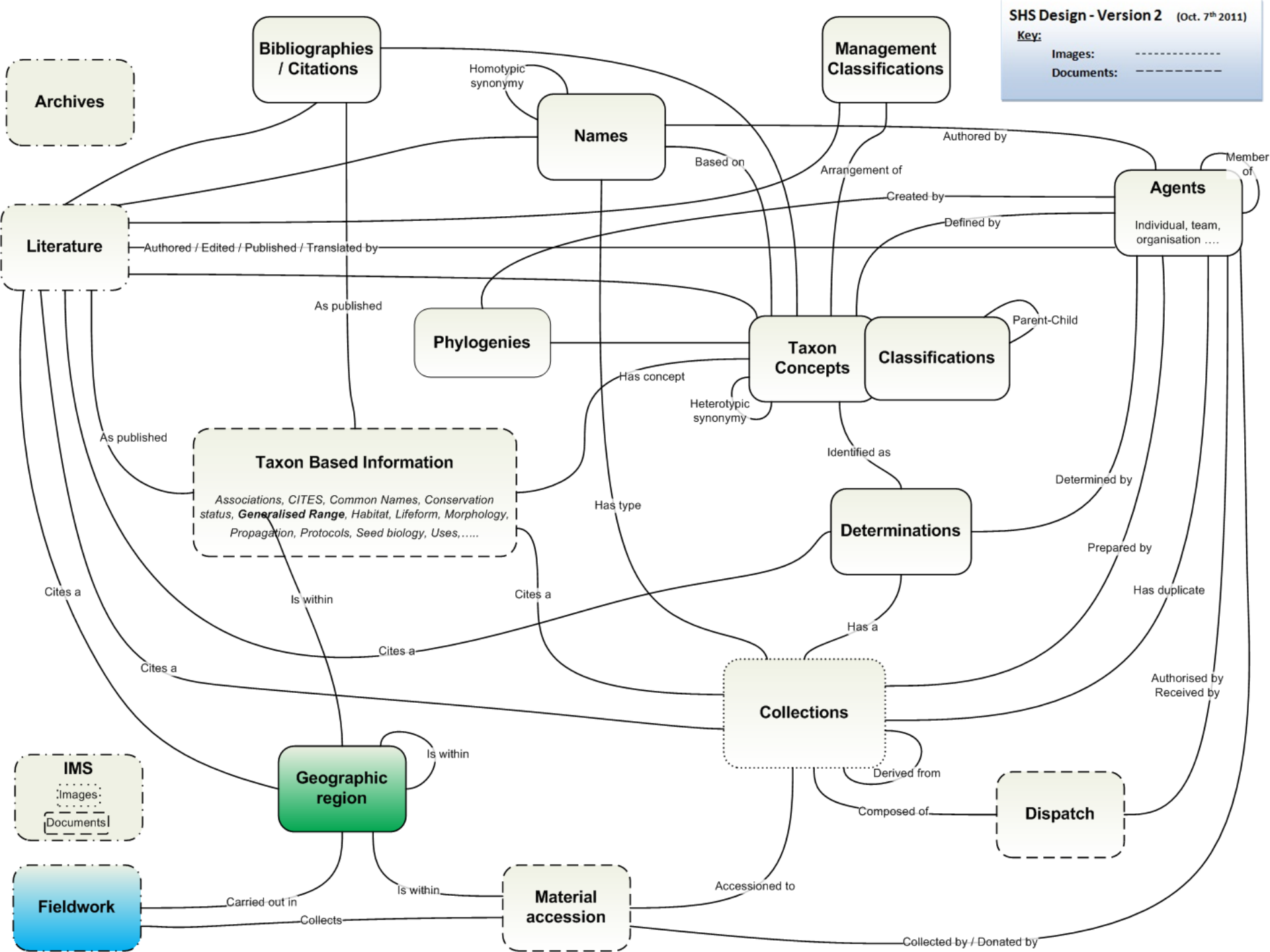
SPNHC 2014
Cardiff, Wales, U.K.

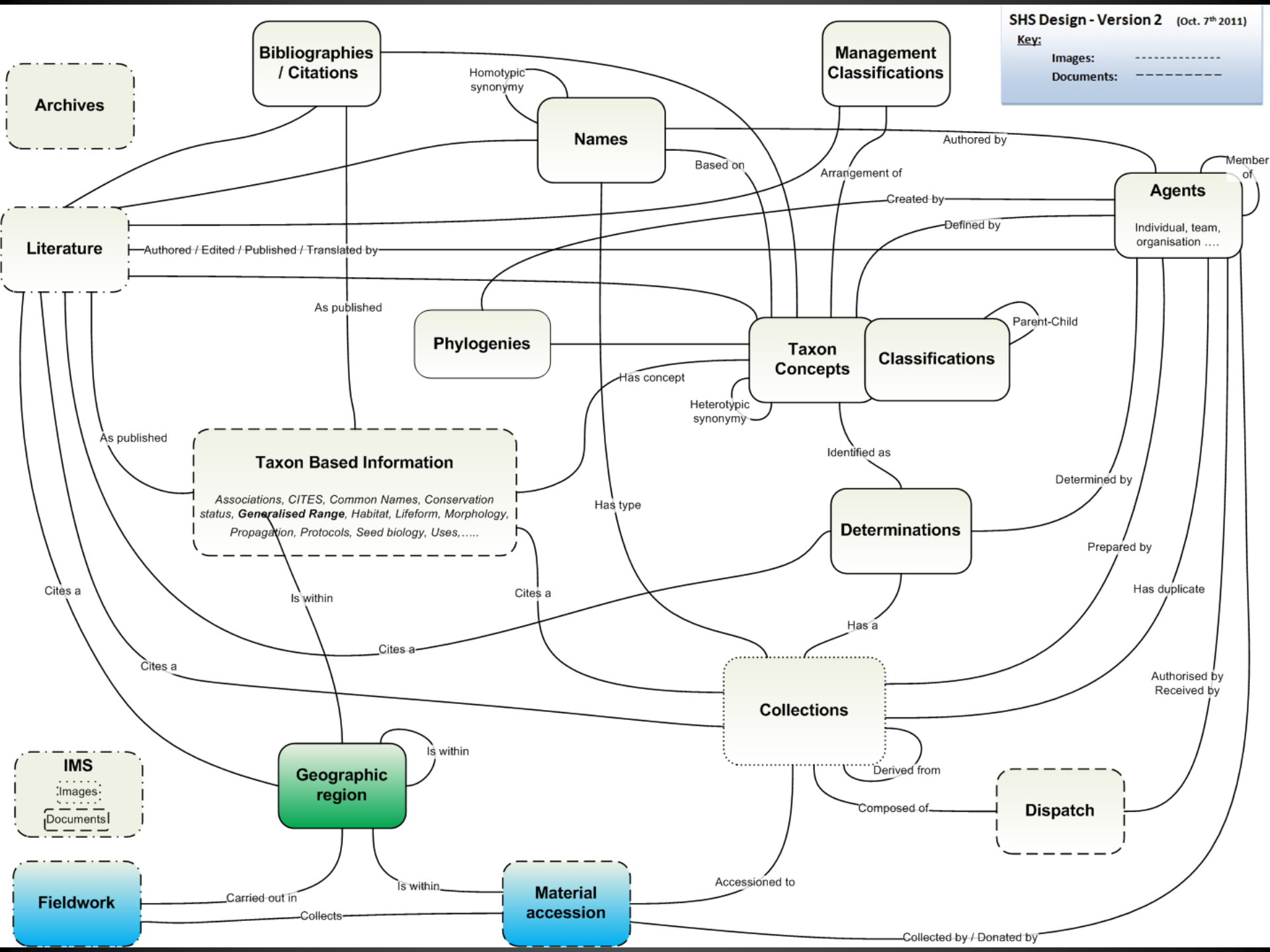


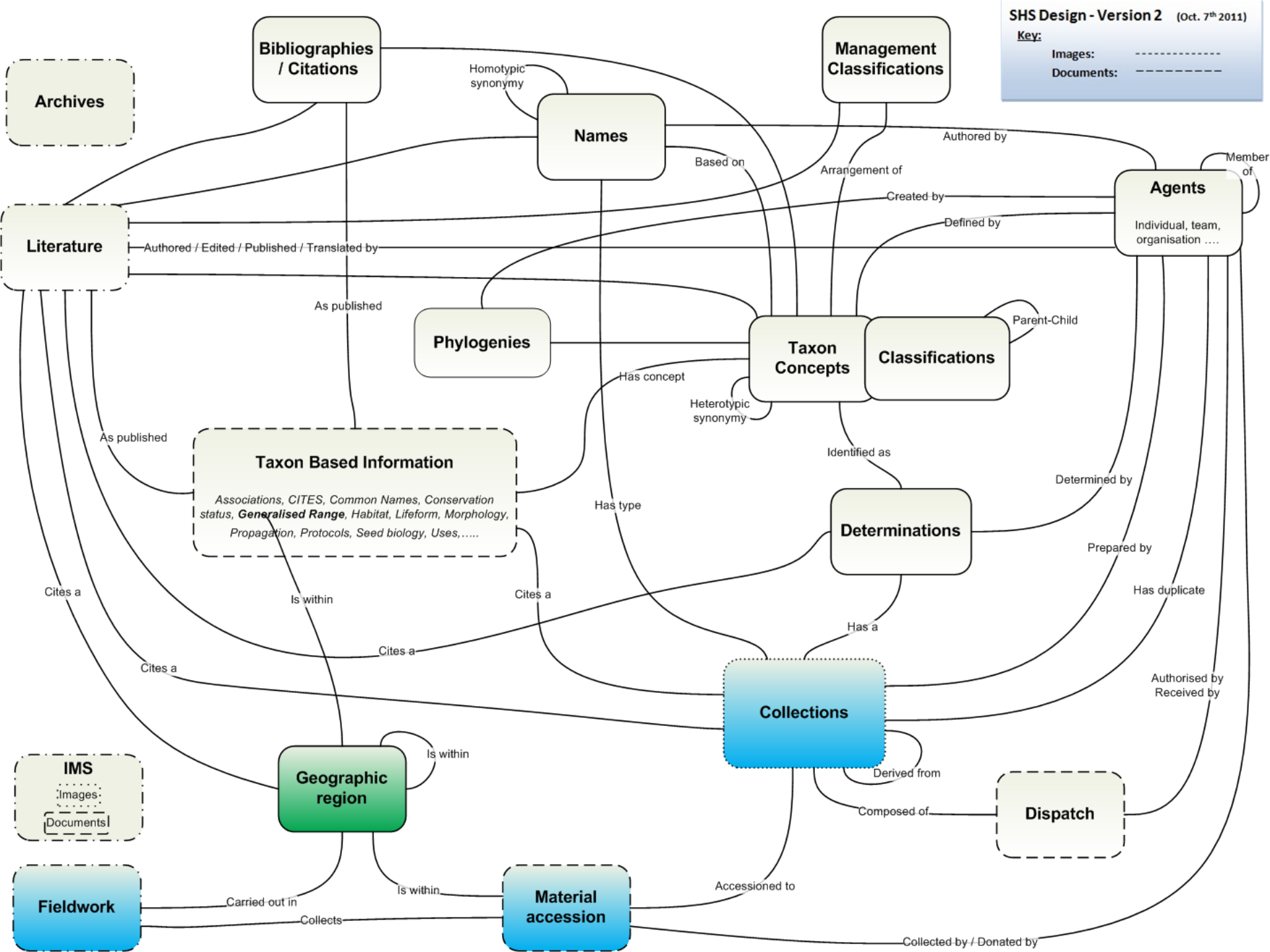


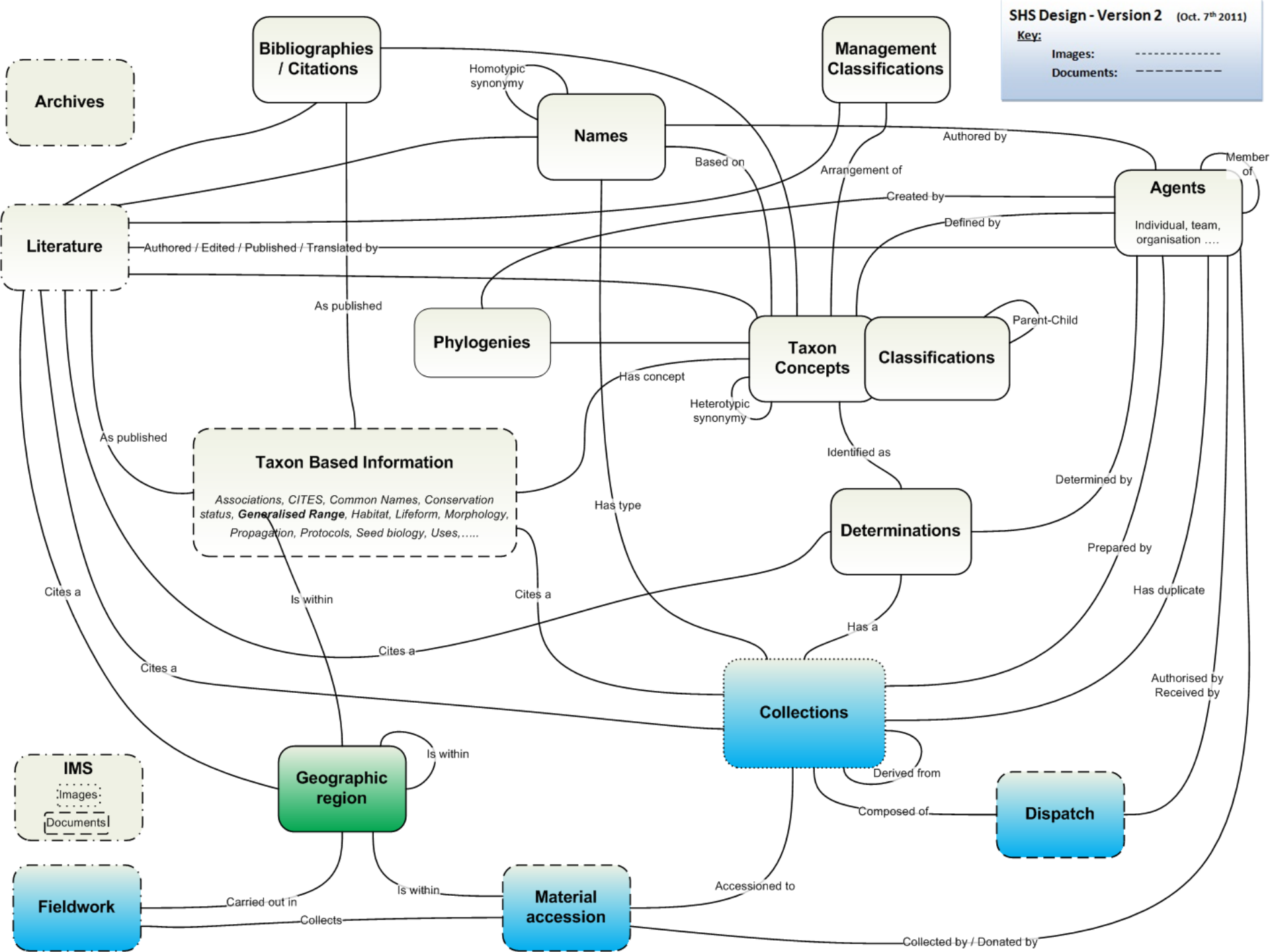


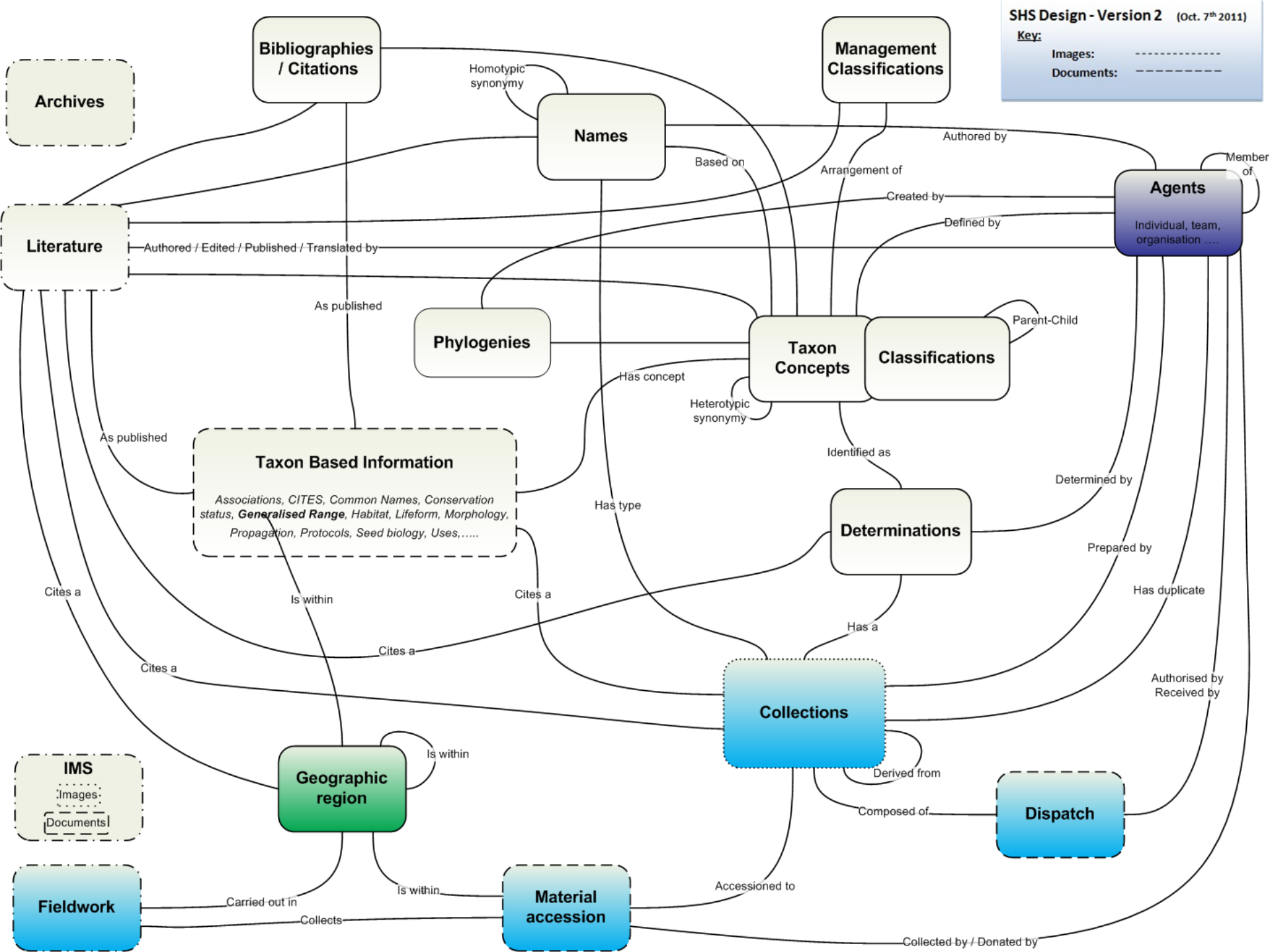


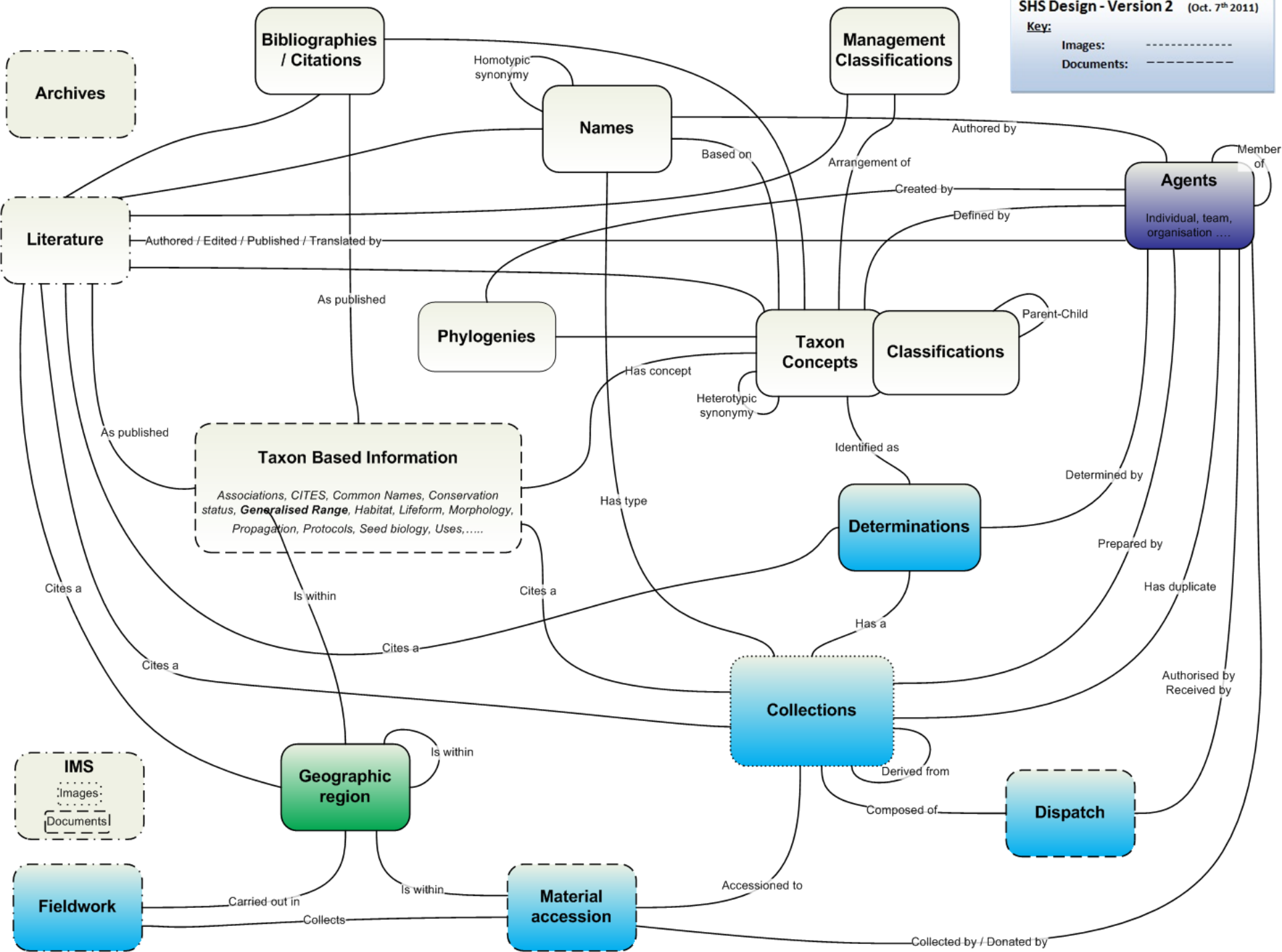


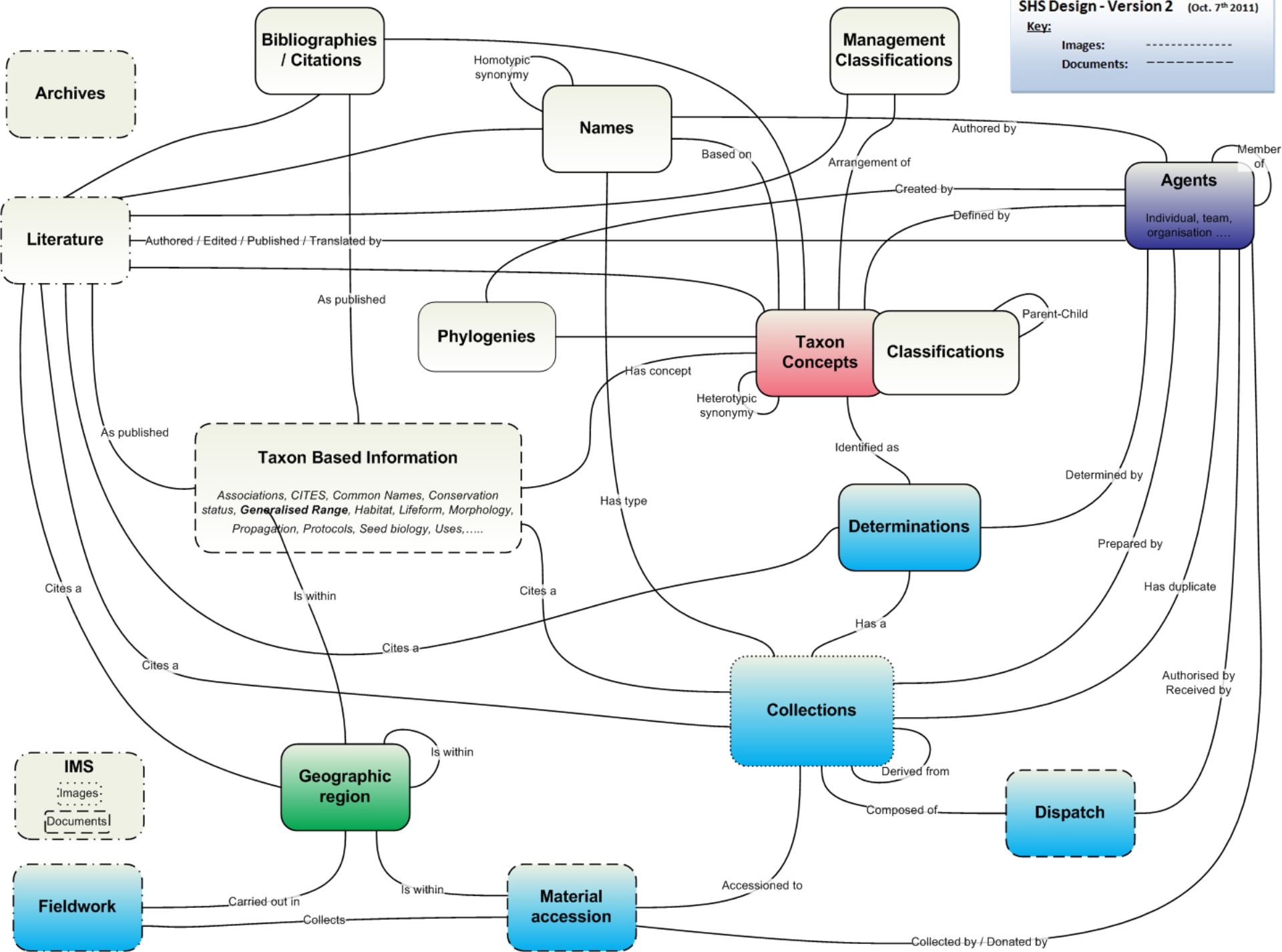


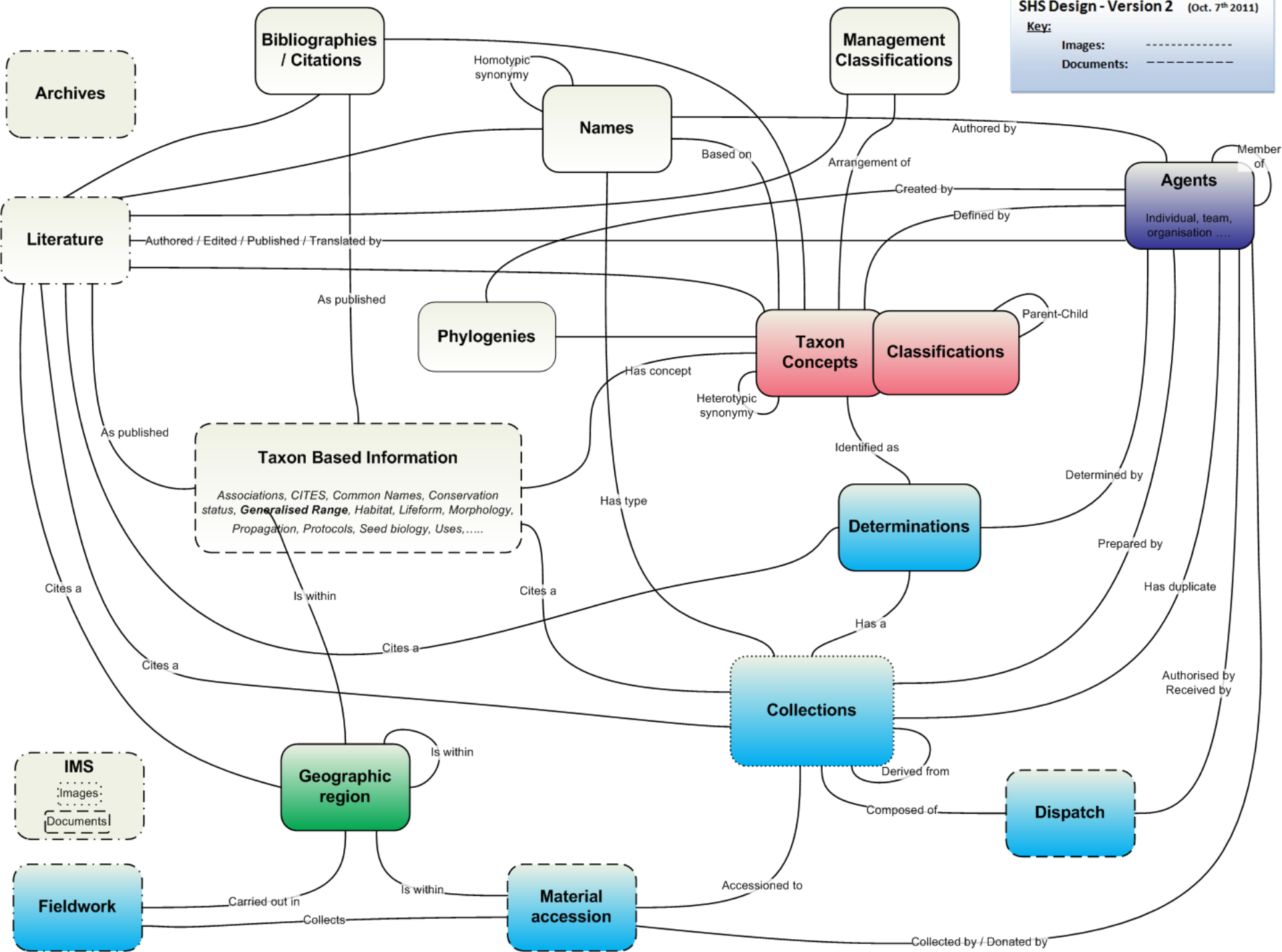


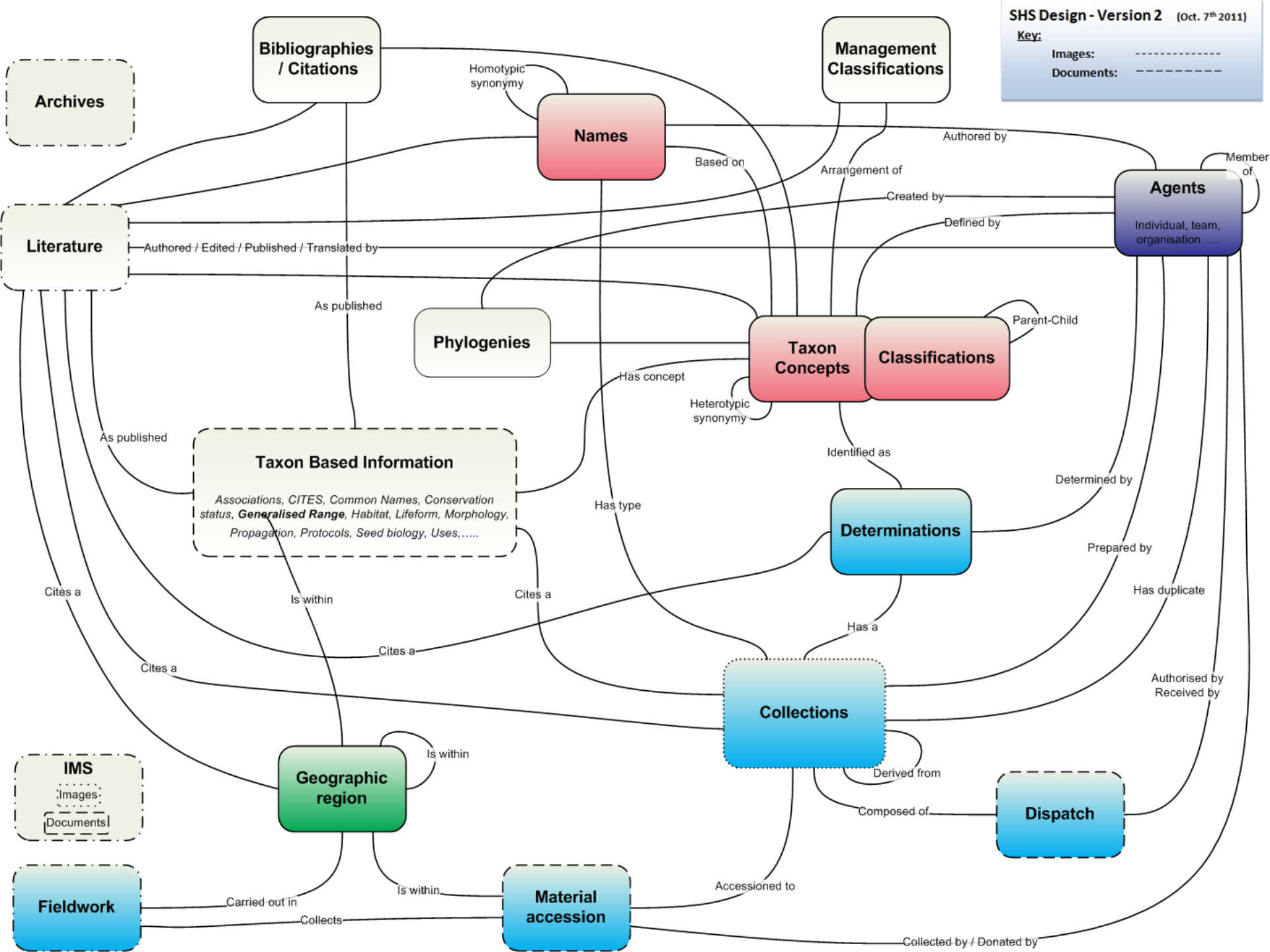


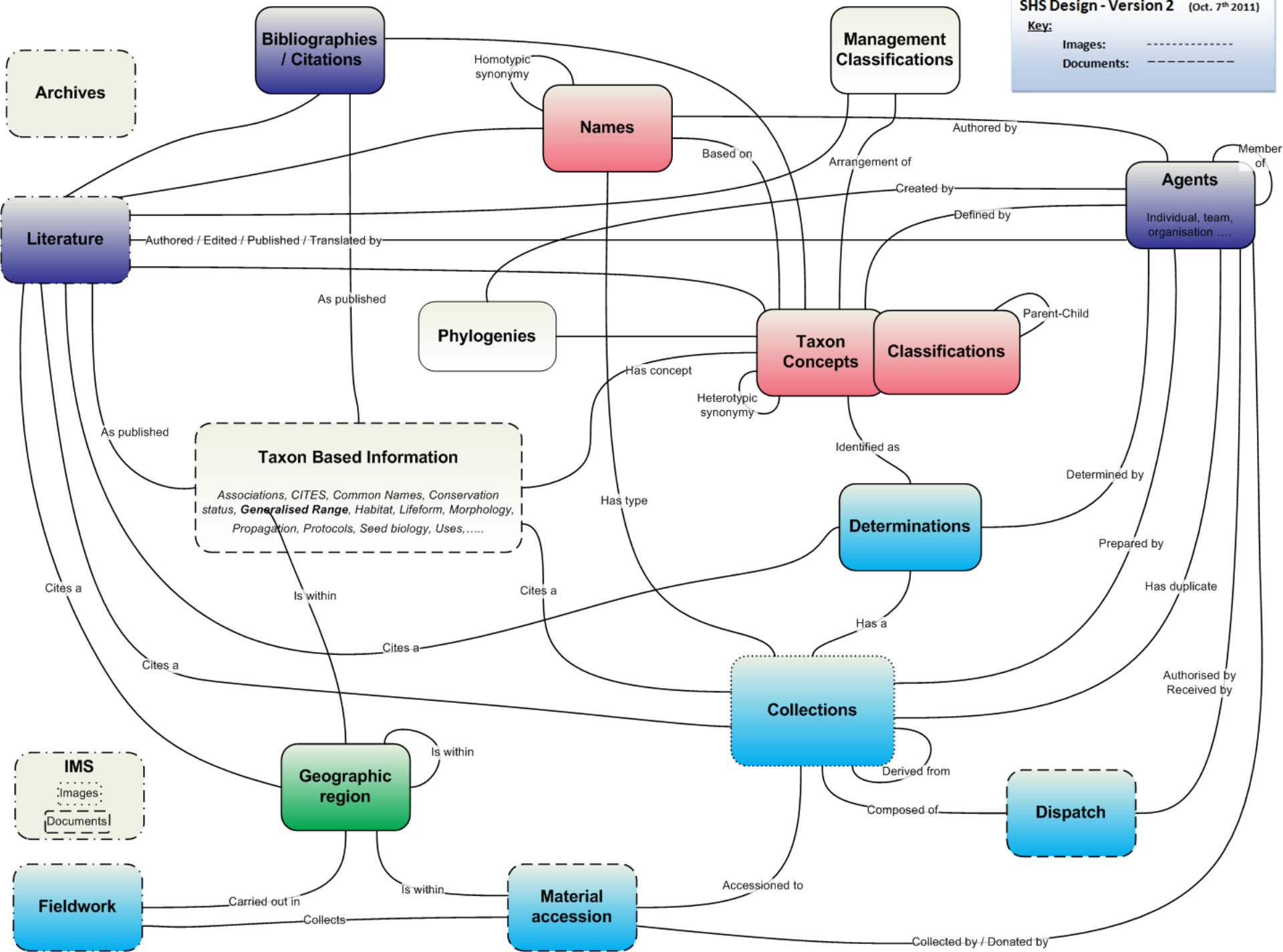


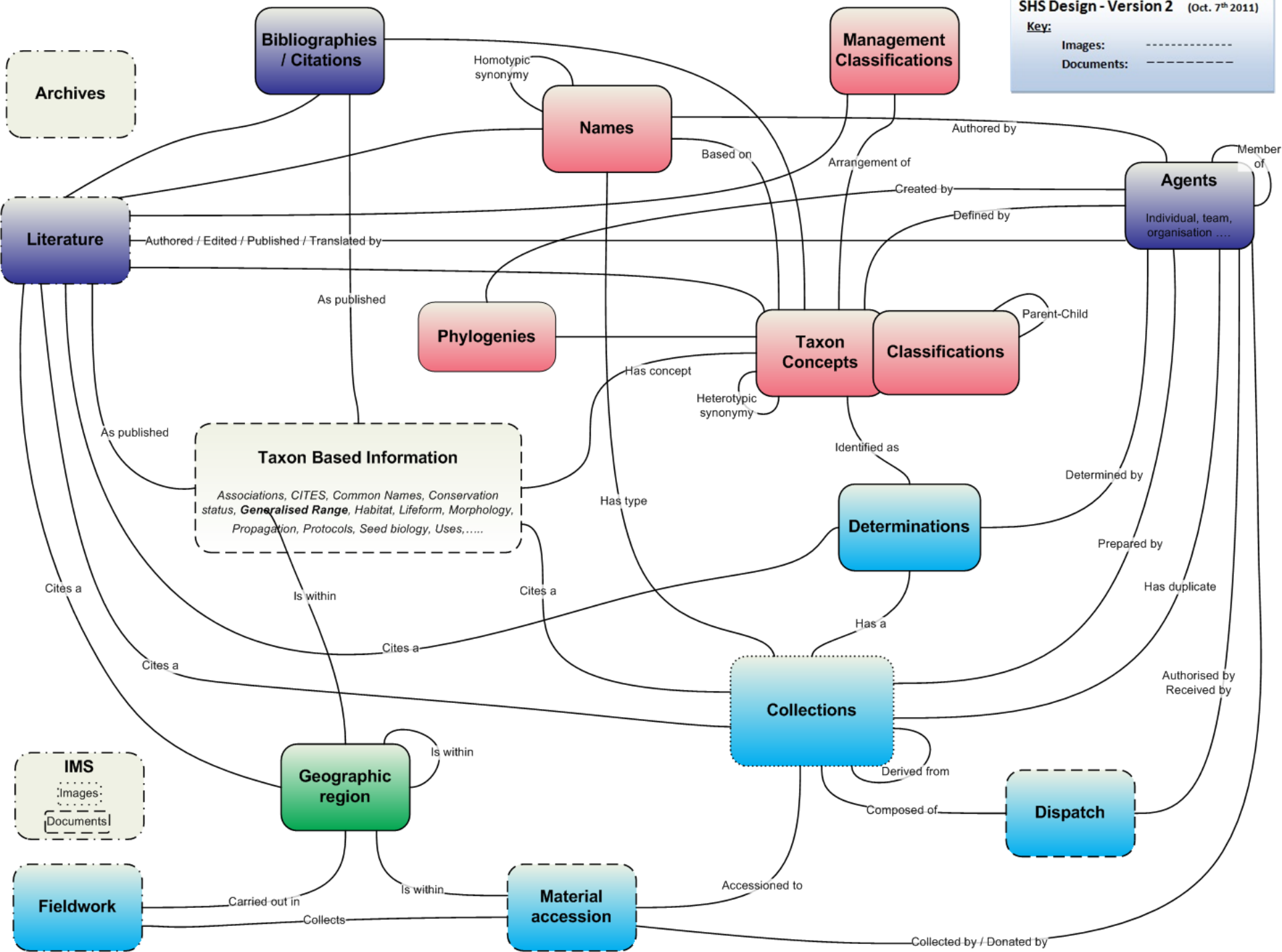


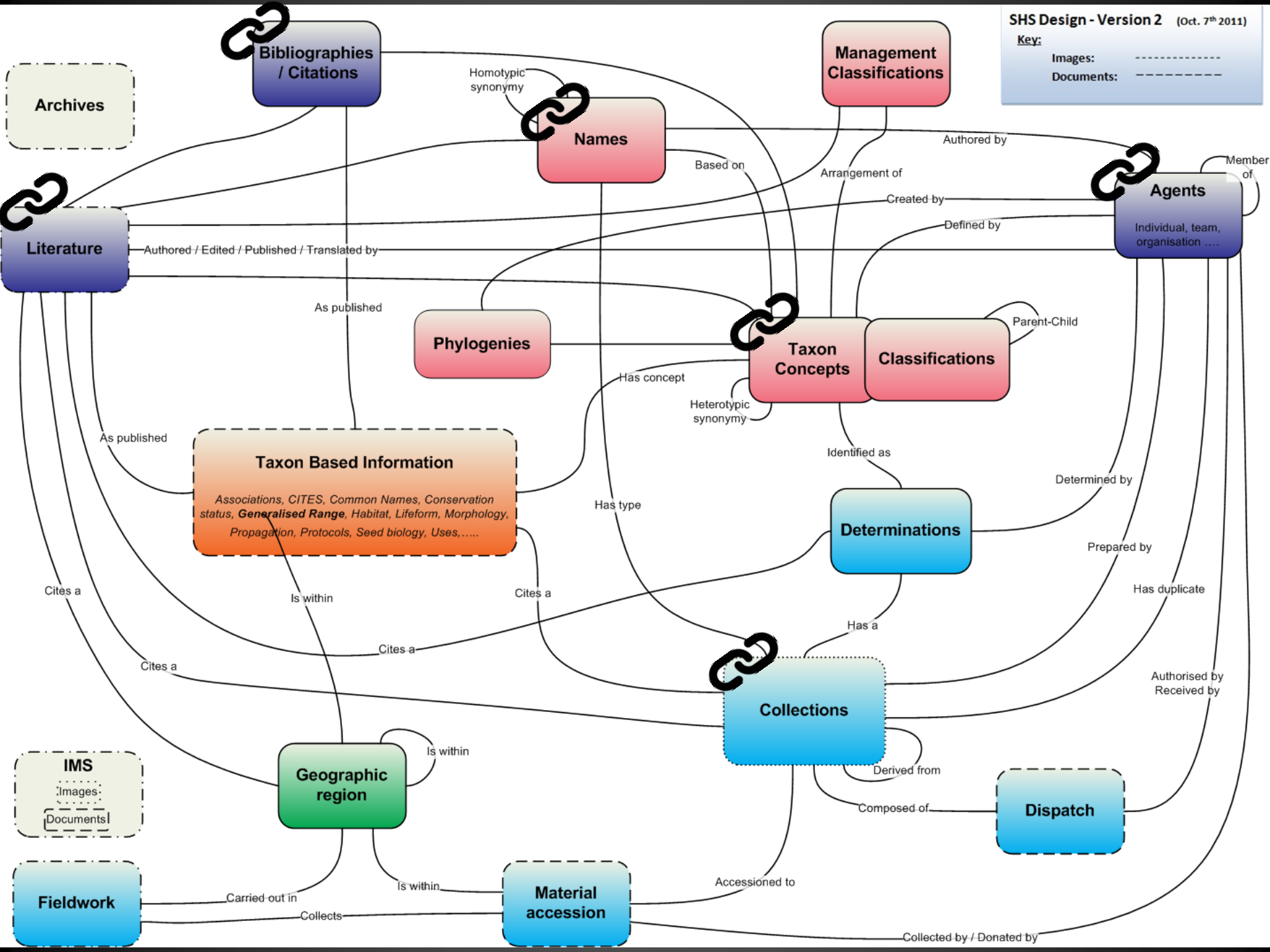












Likelihood of Duplication

Rate of Data Acquisition

Complementarity

Likelihood of Duplication

Rate of Data Acquisition

Complementarity

Thematic Network:

ENBI



European Network for Biodiversity
Information

Objective: To manage an open network of relevant biodiversity information centres in Europe and other countries of the western European palearctic region.

Task 1fM: To provide multi-lingual access to biodiversity information in the Internet

The work plan is focused on user needs, and on making European biodiversity information available for the end-users. The users include government agencies, decision makers, legislators, scientists, companies, and citizens. Also non-European users are very dependent on access to European information, because many data in European repositories originate from non-European (often developing) countries. Understanding the needs of all these kinds of users is paramount for the dissemination of biodiversity knowledge resources, and common access, with attention to multilingual access (WP 11), is a key issue

Funded by the European Commission under its Fifth Framework Programme

To develop and implement technical standards or rules of best practice for specimen databases, particularly with regard to digital imaging.

Chania, Crete in January 2005



How can the incidence of duplicate collections in herbaria be leveraged to improve the rate of data acquisition?

Cathy Furlong 2006.

Master's thesis on likelihood of
specific instances of duplication of
plant specimens between herbaria.

American University, Washington, DC, USA

Cathy Furlong 2006.

Master's thesis on **likelihood** of
specific instances of duplication of
plant specimens between herbaria.

American University, Washington, DC, USA


Question:

If significant resources were applied to completely digitizing ten selected major herbaria, what would be the incidence of duplication with any other herbarium.

What factor or factors can be used to predict the degree of overlap in collectors across herbaria.

- List of collectors in ten major herbaria
- List of collectors in all herbaria
- Size
- Geographic Specialties
- Collections Scope
 - Local
 - Regional
 - National
 - International
 - International over 1.5 M specimens

- List of **collectors** in ten major herbaria
- List of **collectors** in all herbaria
- Size
- Geographic Specialties
- Collections Scope
 - Local
 - Regional
 - National
 - International
 - International over 1.5 M specimens

- List of **collectors** in ten major herbaria
 - List of **collectors** in all herbaria
 - Size
 - Geographic Specialties
 - **Collections Scope**
 - Local
 - Regional
 - National
 - International
 - International over 1.5 M specimens
- 

Likelihood of Duplication

Rate of Data Acquisition

Complementarity



SGR - Scatter Gather Reconcile
discovers and facilitates the re-use of
computerized information from duplicate
specimens. SGR completes the circle of
enterprise-level integration by enabling
write-back from GBIF to Specify databases.

Download data from GBIF for comparison

Create comparison dataset of minimal records containing collector name, collector number, taxon, date, locality

Run SGR Analysis and receive suggested matches



Welcome



Data



Trees



Reports



Interactions



Statistics



Query



Workbench



SGR



Plugins



Lifemapper



Attachments

Q Th*

Matchers

- Botany
- Coll. Event Matcher
- Duplicates Matcher
- Localities Matcher
- Taxon Matcher
- Create Matcher

Match Results

- SGR Example2 Botany
- Process Data Set

Data Sets

- Botany
- eturyuj
- Example SGR
- Example SGR 2
- Herbdata
- Images
- SGR Example2

StartDate: 1996-09-29

LocalityName:

Latitude1:

Longitude1:

Genus: Sicyos

Species: laciniatus

Subspecies1:

County:

Country:

Field Number: 5161

Collector First Name1: E

Collector Last Name1: Carranza

State:

Min Elevation:

ID:	126377804-GBIF	236116660-GBIF	235754298-GBIF
Catalog #:	139603	139603	K423167
Collector/Field #:	5161	5161	5162
Collectors:	E. Carranza G. y C. González	E. Carranza G. y C. González	E. Carranza G.
Taxon Name:	Sicyos laciniatus L.	Sicyos laciniatus L.	Eruca sativa Mill.
Determiner:			L. Hernández
Det. Date:			
Date:	1996-9-29	1996-9-29	1996-9-29
Latitude:			21.633
Longitude:			-101.467
Locality:	OCAMPO	OCAMPO	2 km al NW de Ocar
Municipality:			
County:	OCAMPO	OCAMPO	OCAMPO
State:	GUANAJUATO	GUANAJUATO	Guanajuato
Country:	MEXICO	MEXICO	MEXICO
Institution:	IEB	IEB	IE
Collection:	IEB	IEB	XAL
Source:	GBIF	GBIF	GBIF

2 of 9

☐ Highlight Invalid Cells☐ Highlight New Records

Botany



Save

Grid



Trash

Welcome

SGR Example2

Benefits:

Validation

Augmentation

Considerations:

Time to process SGR vs. direct data entry

Accuracy of suggestions and unmatched data

Likelihood of Duplication

Rate of Data Acquisition

Complementarity



Taxonomic Literature II

[TL-2 Home](#)
[Search](#)
[Read Online](#)
[History](#)
[Acknowledgements](#)
[Downloads](#)
[Help](#)
[Back to Digital Library](#)

.: Taxonomic Literature II (TL-2)

Taxonomic Literature: A selective guide to botanical publications and collections with dates, commentaries and types (Stafleu et al.).

TL-2 is the premier publication of the International Association for Plant Taxonomy (IAPT) and this online version was made possible by the generous cooperation of the IAPT.

In its print form, TL-2 is a 15 volume guide to the literature of systematic botany published between 1753 and 1940. It is organized by author and includes numbered entries for the author's publications. Suggested abbreviations for use in taxonomic publications are provided: abbreviations for the author's name, short titles and abbreviations of the short titles for publications. TL-2 is the standard by which authors' names and titles should be abbreviated.

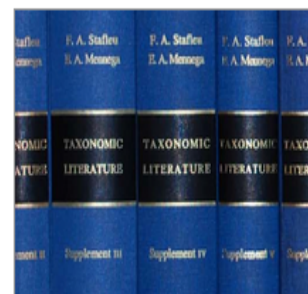


TL-2 at the Smithsonian Institution Libraries

Currently, we offer a basic database searchable by keyword, author name, title number, author name abbreviation, or title abbreviation. We display the search results with the scanned page and the parallel OCR'd and corrected text in a "page turning" application.

The next round of planned improvements, to be done in conjunction with the redesign of the Smithsonian Libraries' website, is to implement Linked Open Data for the entire TL-2 dataset. This computer-friendly format will give each authors and publications a permanent, authoritative URI on the web. These URIs will contain information in both human-readable form (via HTML) and computer-readable form (via RDF/XML.) A SPARQL endpoint will also be provided for querying the linked data.

Further in the future, we wish to perform additional parsing of the data to, for example, extract the Herbaria that contain specimens collected by the authors, link authors and publications to other sources on the web such as the [Virtual International Authority File](#), the [Biodiversity Heritage Library](#),





Elsbeth Haston @emhaston · Mar 4

@rdmpage Common. Plant collectors often collect 5-10 duplicates, 1+ stays in country, 1 to collector's institute, rest to other herbaria

Details

↩ Reply ↻ Retweet ★ Favorite ... More



nicky nicolson

@nickynicolson

+ Follow

@emhaston @rdmpage we've tools to assess and link duplicates, will be demo-ing (and destruction testing) at upcoming Leiden data hack

↩ Reply ↻ Retweet ★ Favorite ... More

RETWEET

1



1:45 AM - 4 Mar 2014

Don't miss any updates from **nicky nicolson**

Full name

Email

Password

Kew at pro-iBiosphere data hackathon

Nicky Nicolson, Matt Blissett

RBG Kew Biodiversity Informatics team



Leiden Data-Hack

- Background on the problem (it's one we all share)
- Pre-existing toolkit
- Activities in the data-hack week
- (Aside – links that already exist in data)
- Conclusion & where next

Shared problem

- We (collections-based systematic research organisations) recognise the same entities
- (Especially in botany and mycology) we have:
 - Lots of data digitally available
 - Culture of referencing authoritative sources (e.g. IPNI / IF)
 - Culture of sharing physical collection objects (specimen duplicates)

Tackling the problem

- To make the most of these resources we need to make links between them.
- Ambitious data integration projects like “World Flora Online”

Data linking tool

- Technology: Java, Spring framework, Lucene
- Rules based
- Armed with a tabular dataset, you:
 - Define zero or more transformers for each field
 - Define how fields must match
 - This is a match configuration.

Examples of transformers

- *Epithet*
mediterraneum → mediterranea
- *NormaliseDiacrits*
Déségl. → Desegl.
- *CleanedPubAuthors*
(L.) A.Gray in Hook.f. → A.Gray
- *SurnameExtractor*
(A.Gray) A.Heller → (Gray) Heller
- *PageExtractor*
37(4): 412 (1977) → 412

Examples of matchers

- *Exact*

“Poa” = “Poa”

- *CommonTokens – how many tokens shared btw the two inputs*

e.g. CapitalLetters

in Beitr. Aethiop. → B A

Beitr. Fl. Aethiop. → B F A = 0.67 ratio

- *Number: 1 = 1*
- *Levenshtein edit distance:*

Plectranthus → Plectranthus (LD1)

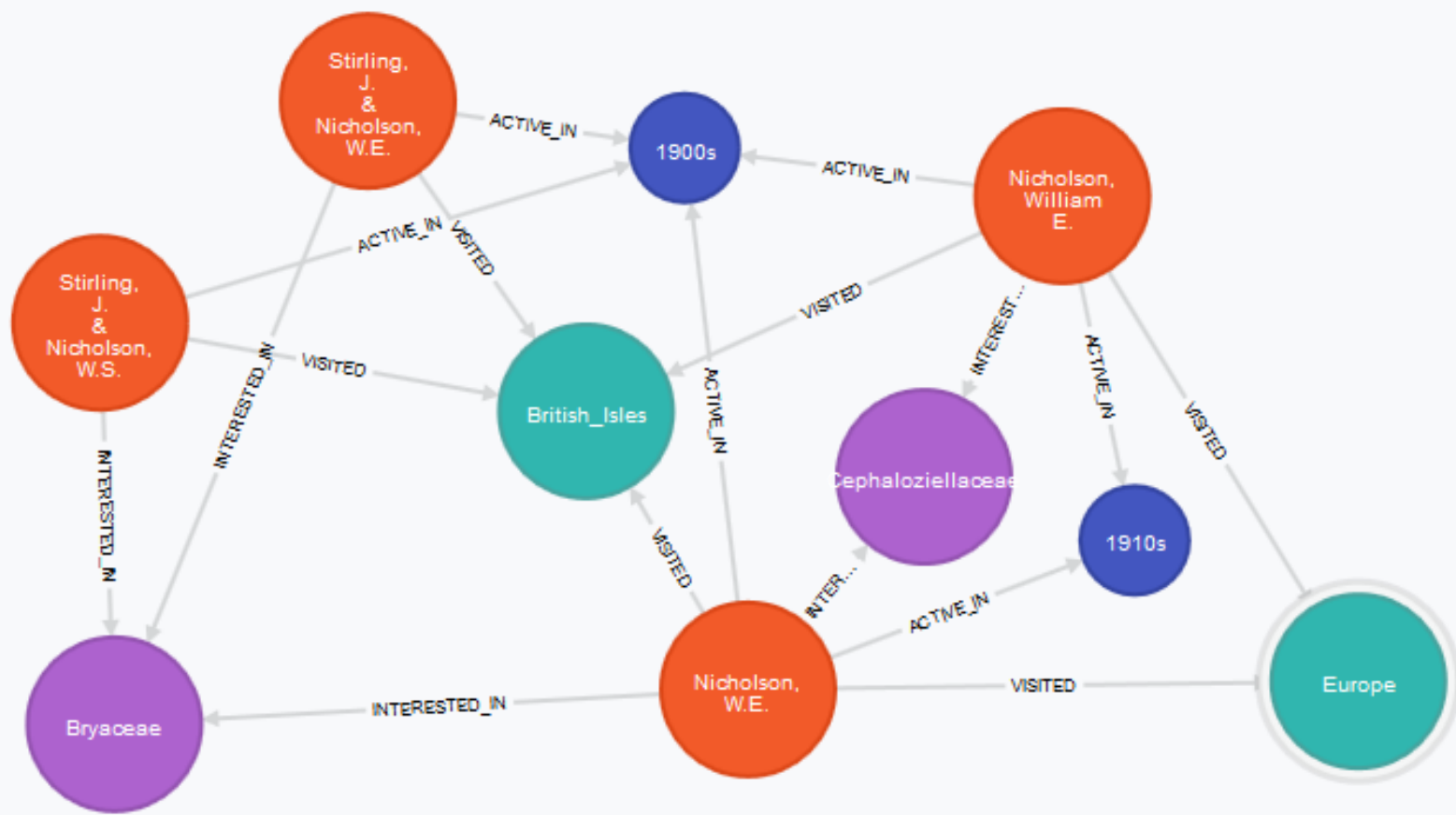
Using the matcher

- A configured match can run against any tabular dataset.
- Accessible as:
 - JSON web service
 - Google Refine reconciliation service (work in progress)
- Transformers can be dropped into Google Refine

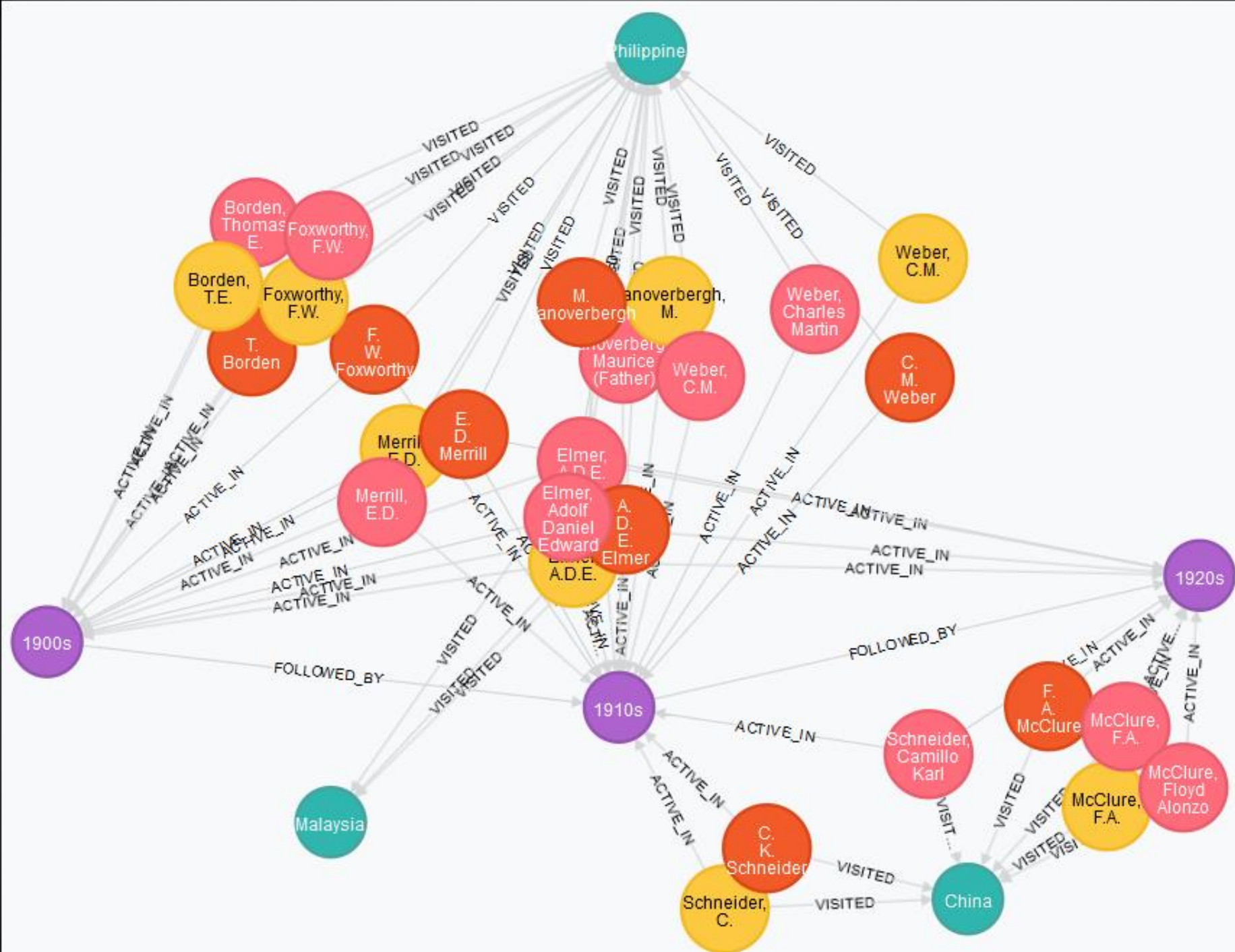
In the data-hack week...

- Worked with the pre-existing toolkit
- Assessed herbaria for duplicates
- Assessed specimens cited in Pensoft journals to link them into digital specimens from herbarium catalogues.
- The work made it obvious that the toolkit would be of use

- Collector
- Place
- Family
- Year



✓ Displaying 10 nodes, 17 relationships



Since the data-hack week...

- Further used the toolkit in the assembly of a proof of concept “World Flora Online”
- We plan to open-source it to allow data custodians to “self serve” in linking their data.
- Longer term we could connect a lot of content, and share digitisation / standardisation effort as a result of this linking.



Plants 2020

Supporting the implementation of the
Global Strategy for Plant Conservation

[Home](#) > The World Flora Online Project



Plants 2020

A GSPC toolkit

The Global Partnership for
Plant Conservation

World Flora Online Project

Memorandum of
Understanding

Consortium members

WFO Council meetings

The World Flora Online Project

Target 1 of the GSPC calls for "A widely accessible working list of known plant species as a step towards a complete world flora"

The World Flora Online Project has been established in response to Target 1 of the GSPC.

The terms and technical rationale for Target 1 suggest that the Flora should include accepted names and a comprehensive synonymy, building on the results of the previous objectives for Target 1 (dated 2002 - 2010), aimed to develop "a widely accessible working list of known plant species as a step towards a complete world flora." New knowledge should also be incorporated as it becomes available. Target 1 of the first phase of the GSPC was achieved at the end of 2010, through The Plant List (www.theplantlist.org).

Establishment of the World Flora Online project

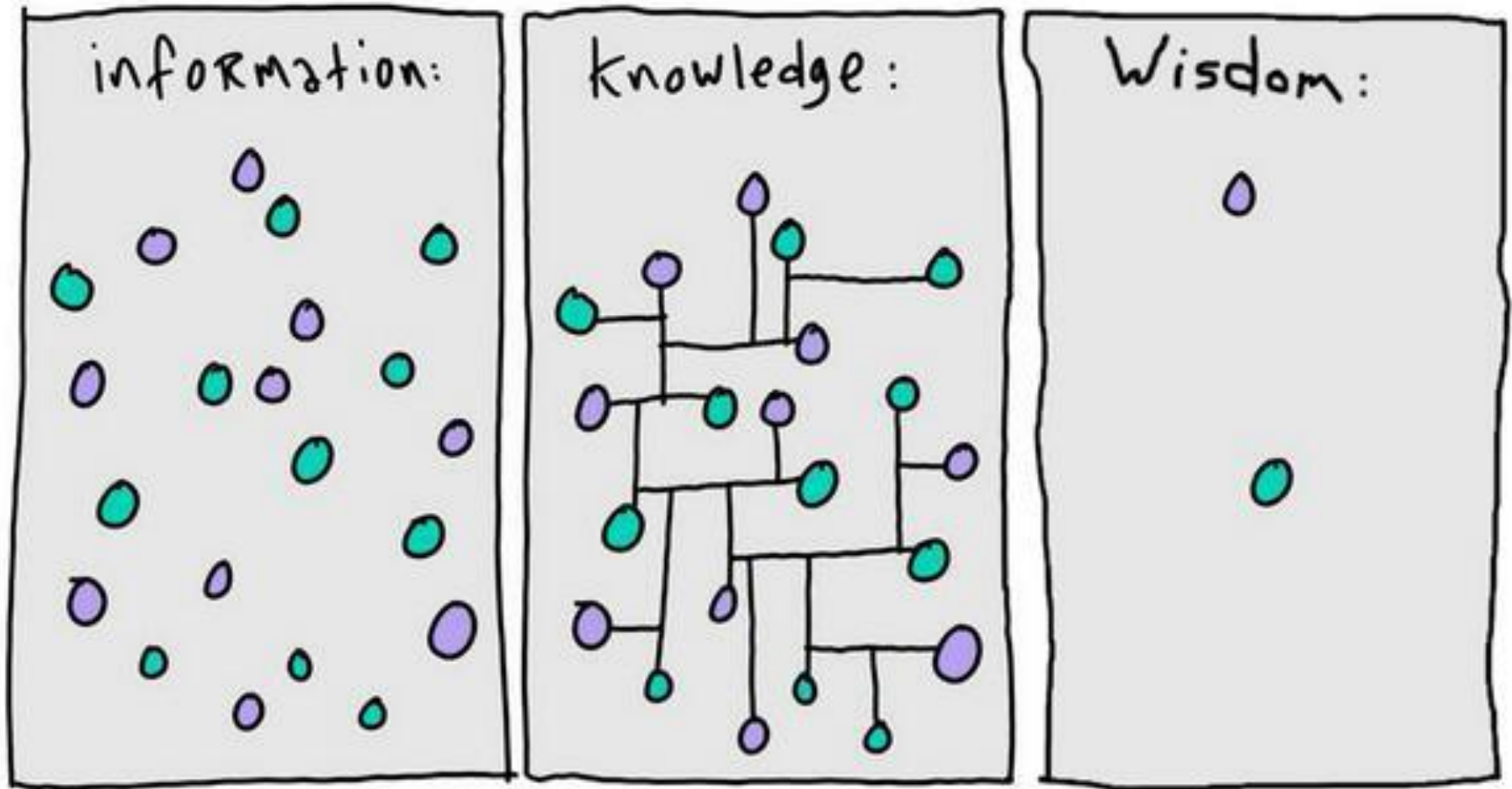
- An initial MOU between Missouri Botanical Garden, Royal Botanic Gardens, Kew, Royal Botanic Garden Edinburgh and New York Botanical Garden was signed February 29, 2012.
- The World Flora Online was launched in India, at an event held during the 11th Conference of the Parties to the Convention on Biological Diversity in October, 2012
- In July 2012 the first World Flora Online Meeting was held at Missouri Botanical garden, USA



BGCI

Plants for the Planet





n.nicolson@kew.org / [@nickynicolson](https://twitter.com/nickynicolson)

m.blissett@kew.org

a.paton@kew.org