



GLOBAL
BIODIVERSITY
INFORMATION
FACILITY

Biodiversity
Information
Standards
TDWG

Biodiversity Portals - Implications for TDWG

Donald Hobern - GBIF Deputy Director for Informatics

dhobern@gbif.org

September 2007

TDWG and federated search protocols

- DiGIR/BioCAsE/TAPIR were developed to support federated searches
- The intended use case was as follows:
 1. A user submits a search request
 - *Find occurrences of Chiroptera from before 1950*
 2. A workflow application passes the request to relevant data providers
 3. Each provider responds with at least the first page of matching records
 4. The workflow application returns the combined results to the user (with support for retrieving records not returned on the first request)

Intended use case for TDWG standards - simple

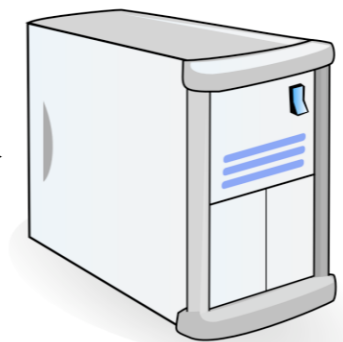


Key	Registry	Protocol
1	Mammal Collection A	DiGIR
2	Mammal Collection B	DiGIR
3	Zoological Museum	DiGIR
4	Botanic Garden	DiGIR

Registry



Portal



Mammal Collection A



Mammal Collection B



Zoological Museum

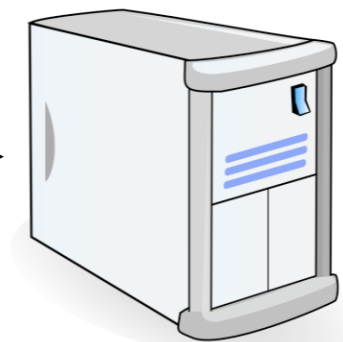


Botanic Garden

Intended use case for TDWG standards - simple



Request: Find Chiroptera
occurrences prior to 1950



Key	Registry	Protocol
1	Mammal Collection A	DiGIR
2	Mammal Collection B	DiGIR
3	Zoological Museum	DiGIR
4	Botanic Garden	DiGIR

Registry



Portal



Mammal Collection A



Mammal Collection B

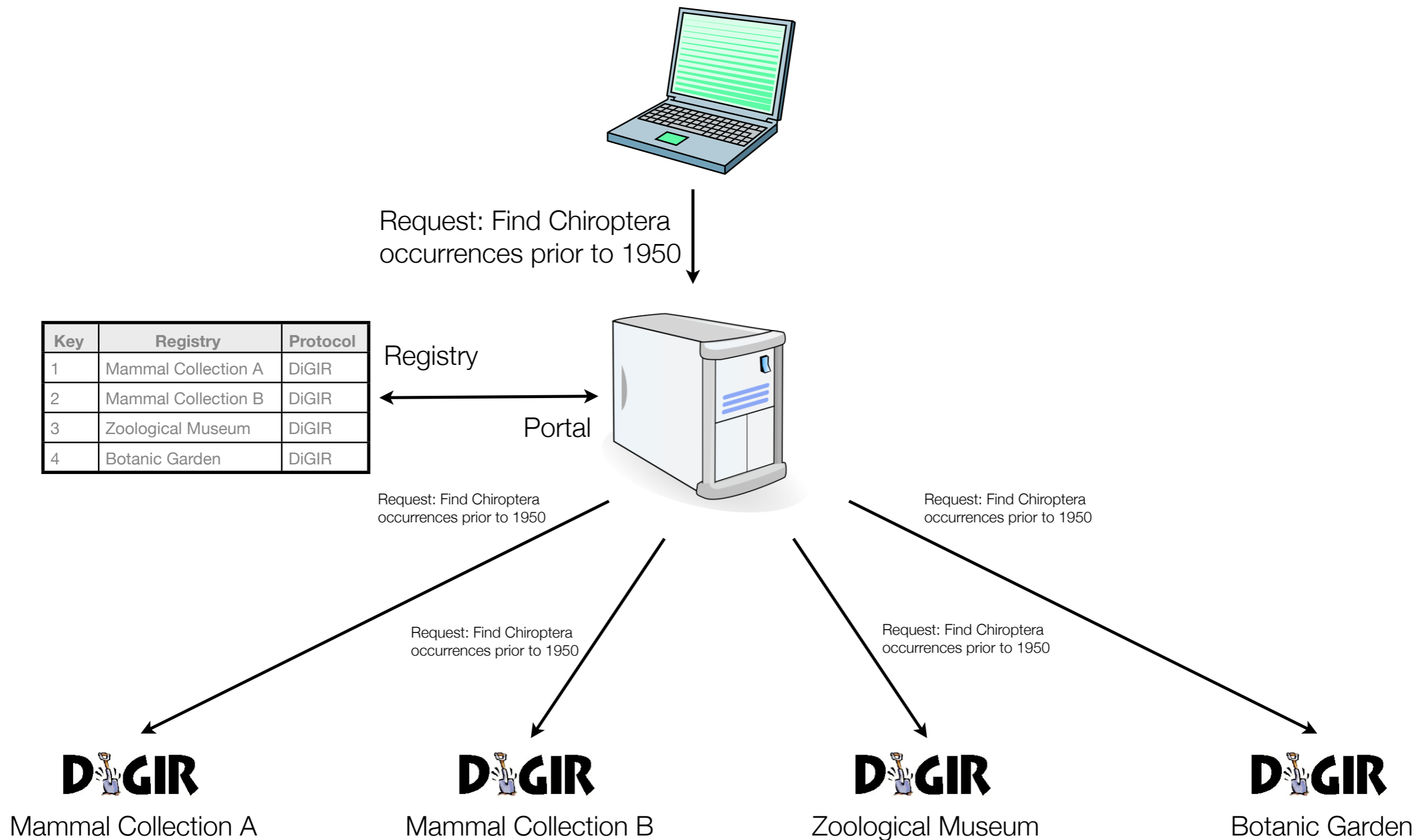


Zoological Museum

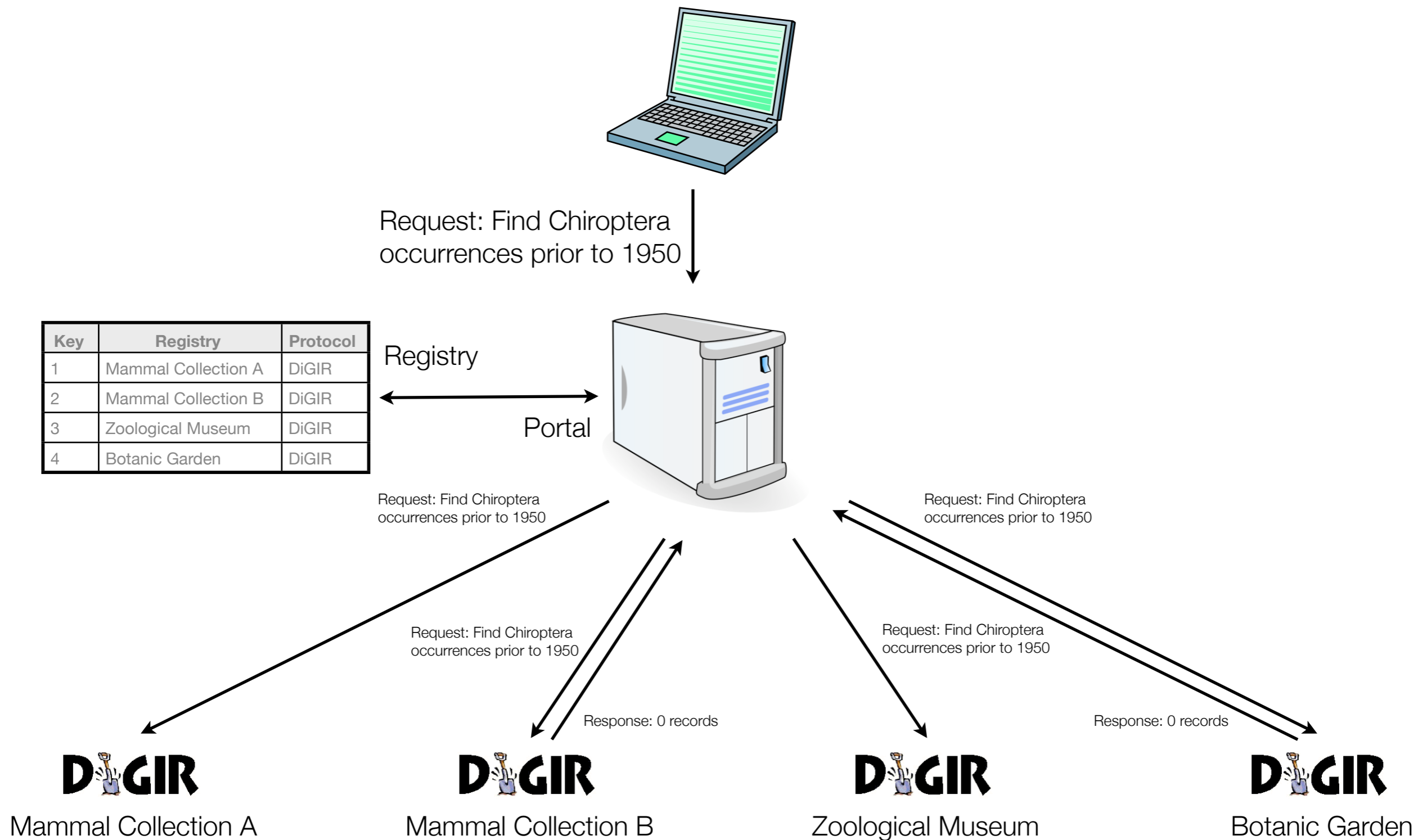


Botanic Garden

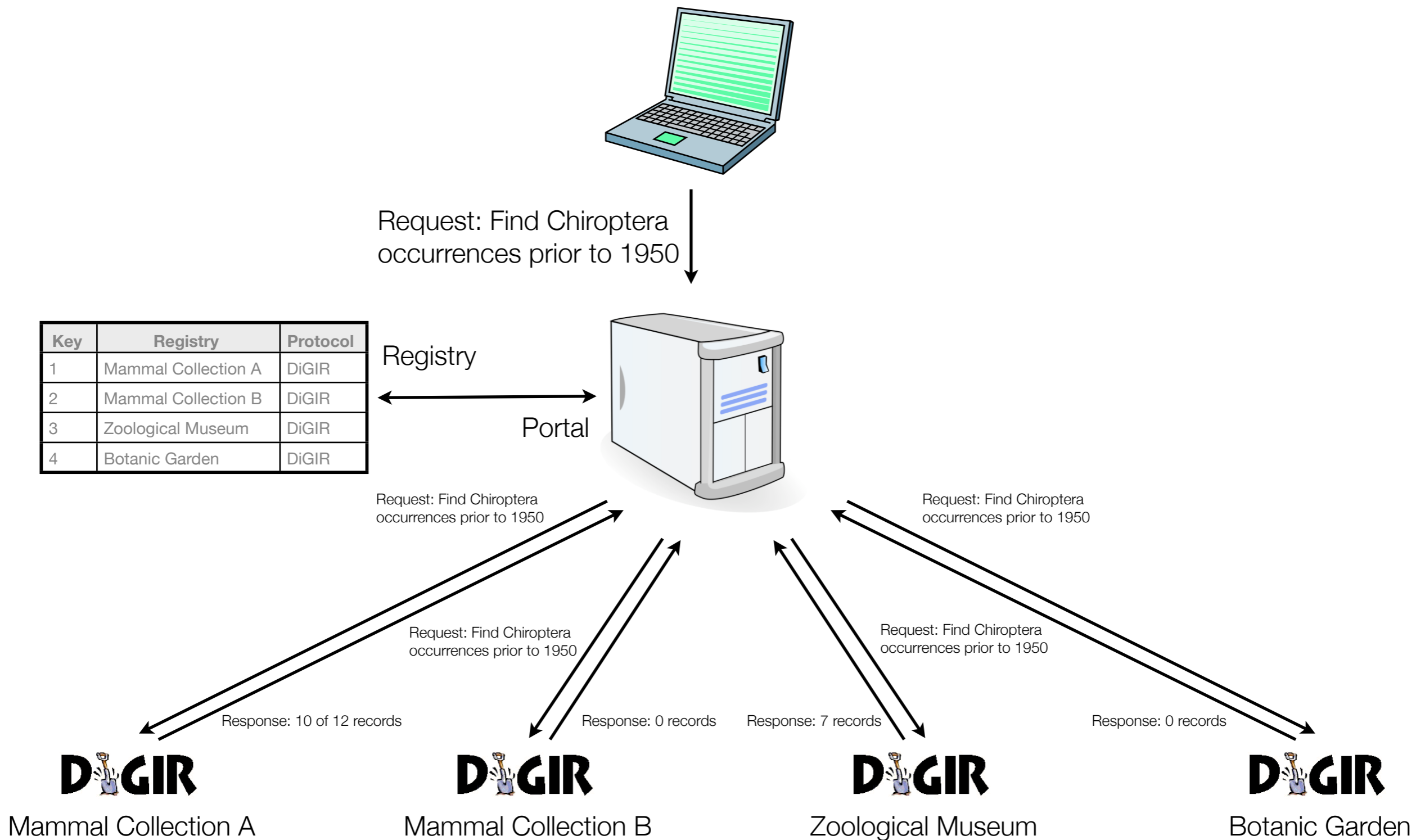
Intended use case for TDWG standards - simple



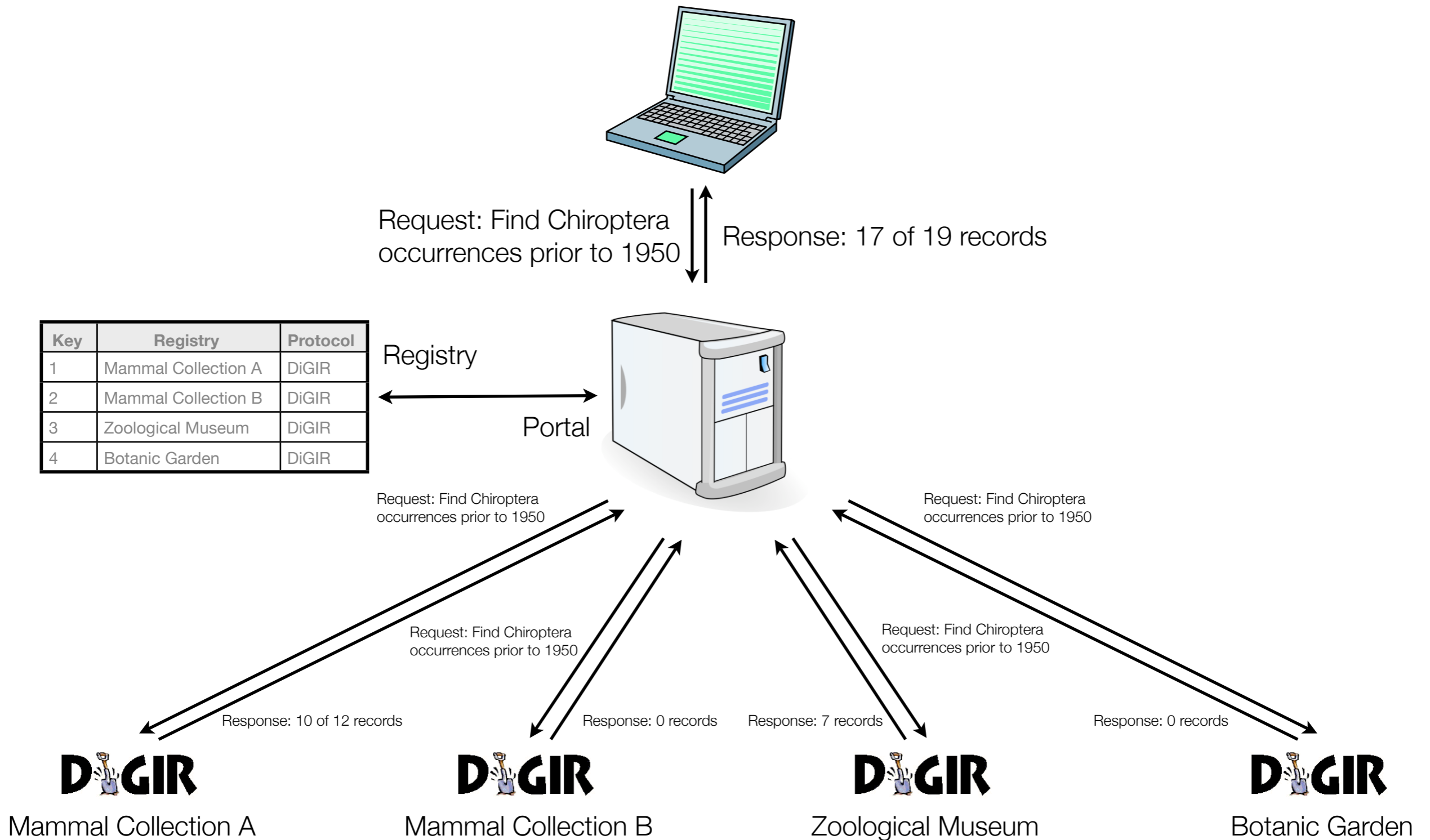
Intended use case for TDWG standards - simple



Intended use case for TDWG standards - simple



Intended use case for TDWG standards - simple



Basic requirements for portal function

- The portal should maintain the following information:
 - Basic technical metadata for each dataset (endpoint, data standards, etc.)
 - Session information to support paging through matching data sets
 - Ideally - knowledge of each dataset's content so requests can be forwarded only to relevant data providers
 - Ideally - domain knowledge to enhance requests, e.g. to use synonyms as well as the accepted name for a species

Intended use case for TDWG standards - improved

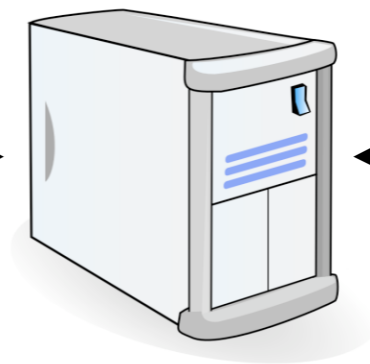


User A

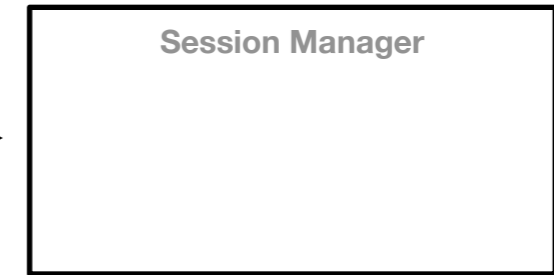
Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Registry

Portal



Session Manager



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

Intended use case for TDWG standards - improved



User A

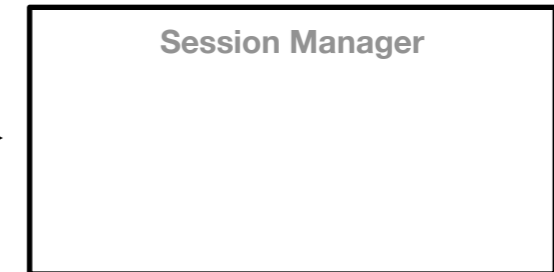
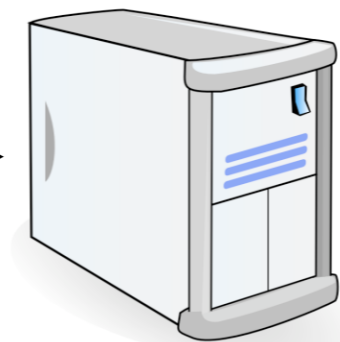
Request: Find Chiroptera
occurrences prior to 1950



Registry

Portal

Session Manager



Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

Intended use case for TDWG standards - improved



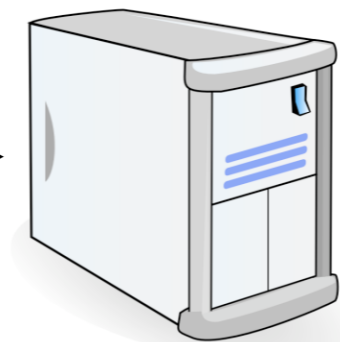
User A

Request: Find Chiroptera
occurrences prior to 1950



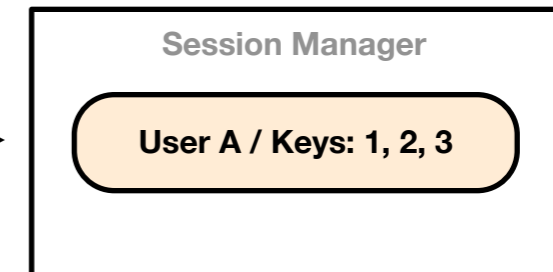
Registry

Portal



Session Manager

User A / Keys: 1, 2, 3



Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

Intended use case for TDWG standards - improved



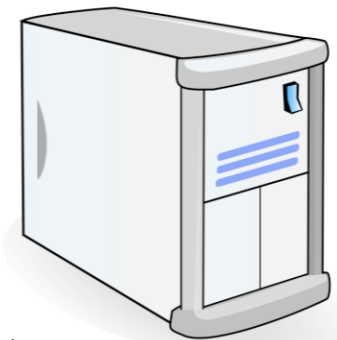
User A

Request: Find Chiroptera occurrences prior to 1950



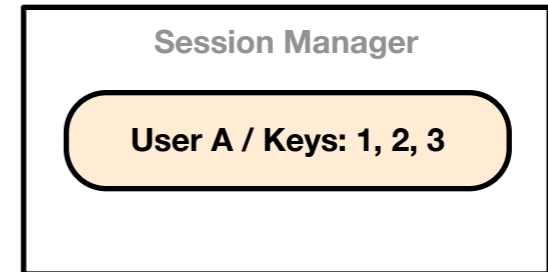
Registry

Portal



Session Manager

User A / Keys: 1, 2, 3



Request: Find Chiroptera occurrences prior to 1950

Request: Find Chiroptera occurrences prior to 1950

Request: Find Chiroptera occurrences prior to 1950



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

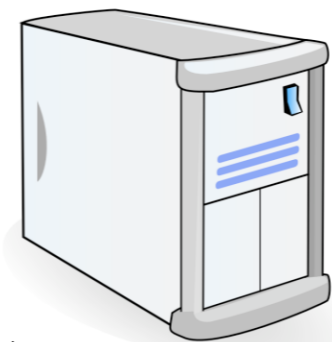
Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Intended use case for TDWG standards - improved



User A

Request: Find Chiroptera occurrences prior to 1950



Registry

Portal

Session Manager

User A / Keys: 1, 2, 3



Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Request: Find Chiroptera occurrences prior to 1950

Request: Find Chiroptera occurrences prior to 1950

Request: Find Chiroptera occurrences prior to 1950

Response: 10 of 12 records

Response: 0 records

Response: 7 records



Mammal Collection A



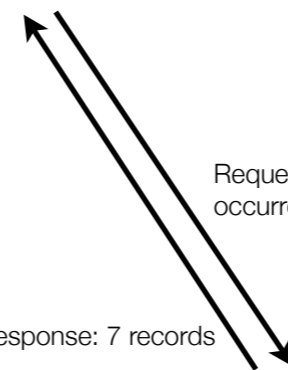
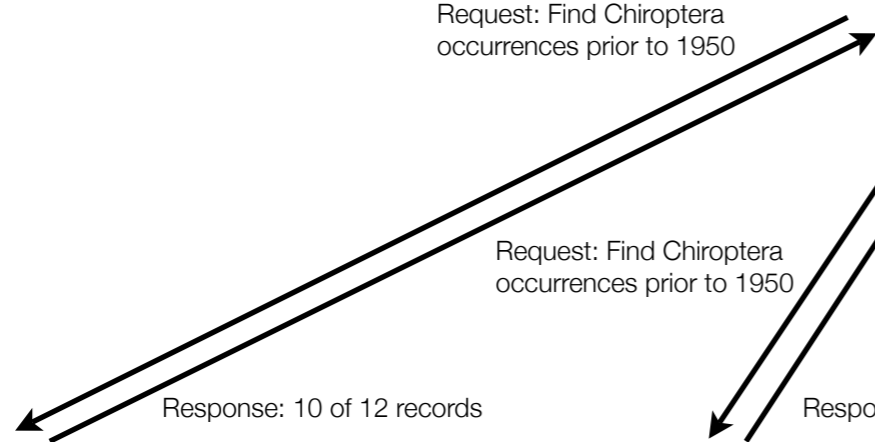
Mammal Collection B



Zoological Museum



Botanic Garden



Intended use case for TDWG standards - improved



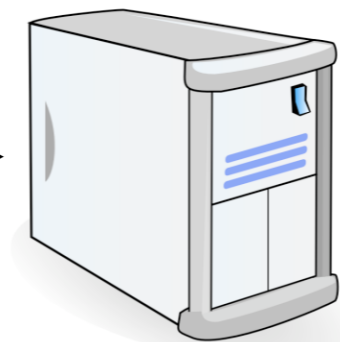
User A

Request: Find Chiroptera
occurrences prior to 1950



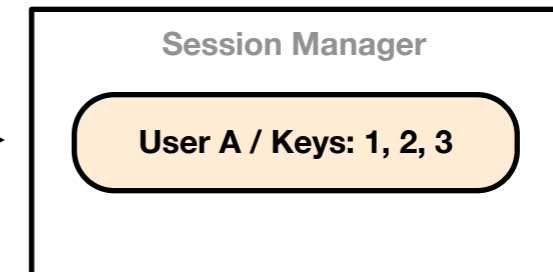
Registry

Portal



Session Manager

User A / Keys: 1, 2, 3



Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae



Mammal Collection A



Mammal Collection B



Zoological Museum



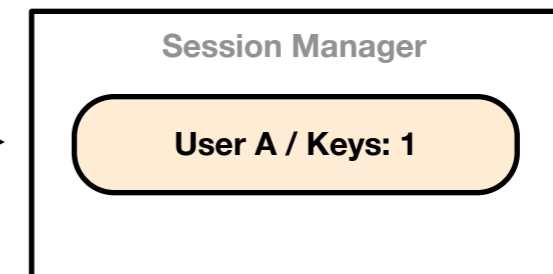
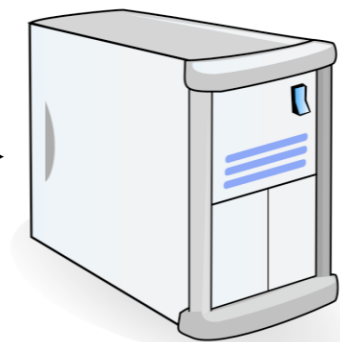
Botanic Garden

Intended use case for TDWG standards - improved



User A

Request: Find Chiroptera
occurrences prior to 1950



Registry

Portal

Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae



Mammal Collection A



Mammal Collection B

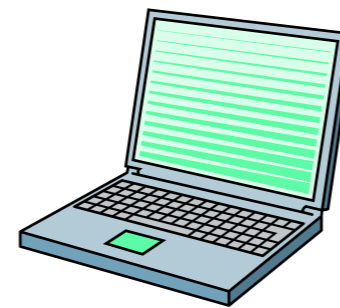


Zoological Museum



Botanic Garden

Intended use case for TDWG standards - improved



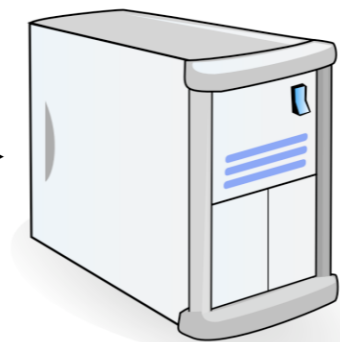
User A

Request: Find Chiroptera occurrences prior to 1950



Registry

Portal



Session Manager

User A / Keys: 1



Request: Find Chiroptera occurrences prior to 1950 from record 11



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Intended use case for TDWG standards - improved



User A

Request: Find Chiroptera
occurrences prior to 1950



Registry

Portal

Session Manager

User A / Keys: 1

Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Request: Find Chiroptera
occurrences prior to 1950
from record 11

Response: 2 of 2 records



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

Intended use case for TDWG standards - improved



User A

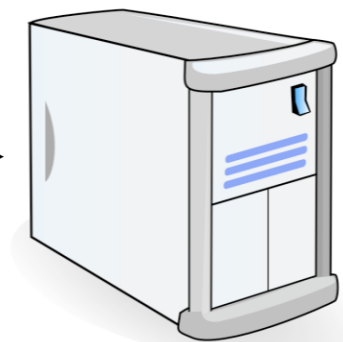
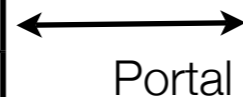
Request: Find Chiroptera occurrences prior to 1950

Response: 19 of 19 records



Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Registry



Session Manager



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

Scalability issues

- There are several problems which appear as network sizes increase:
 - It is wasteful to forward every request to every potentially relevant data provider
 - Many requests are too general for any datasets to be excluded in advance
 - *Find records for any date between 1 Jan 1990 and 31 December 1999*
 - At any time some providers will be off-line
 - Some providers cannot handle complex requests, or respond very slowly
- Most portals (GBIF included) have used an index/cache to address these issues

Use case for TDWG standards - with index/cache

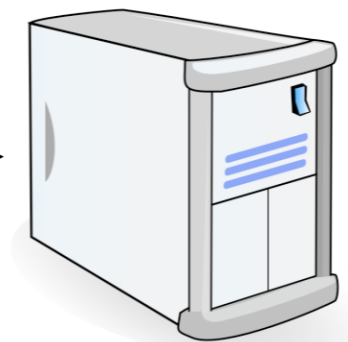


User A

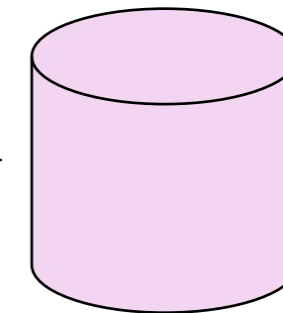
Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Registry

Portal



Index



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

Use case for TDWG standards - with index/cache

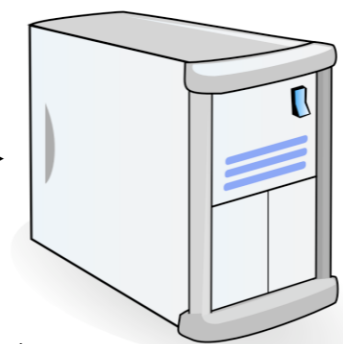


User A

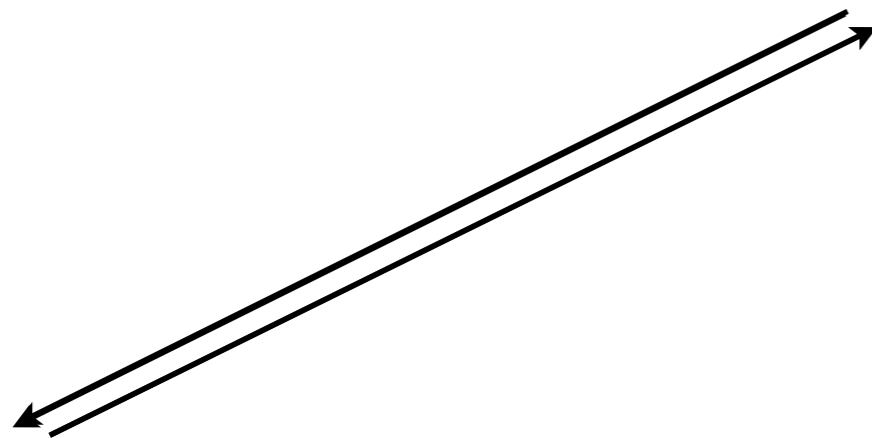
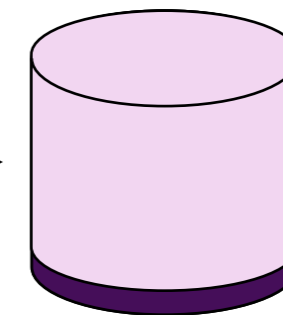
Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Registry

Portal



Index



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

Use case for TDWG standards - with index/cache

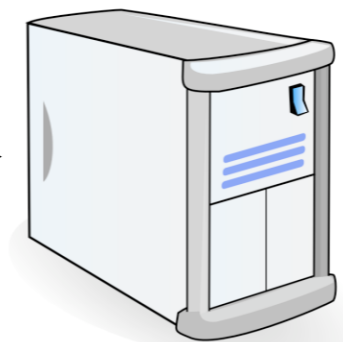


User A

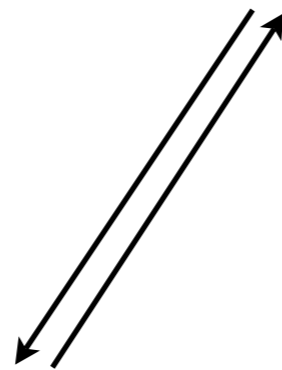
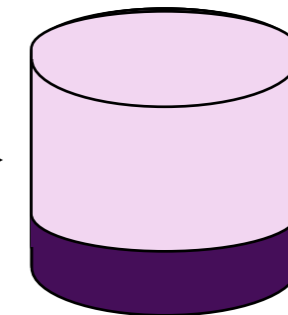
Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Registry

Portal



Index



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

Use case for TDWG standards - with index/cache

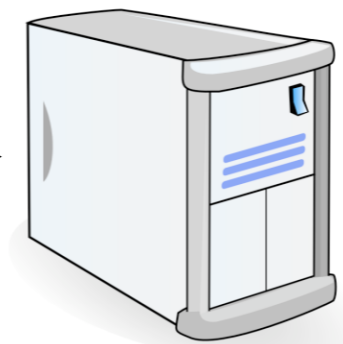


User A

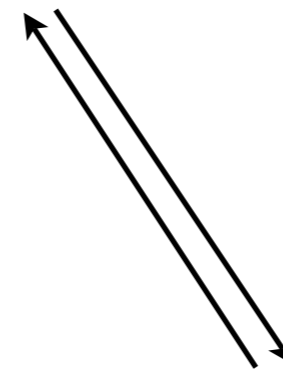
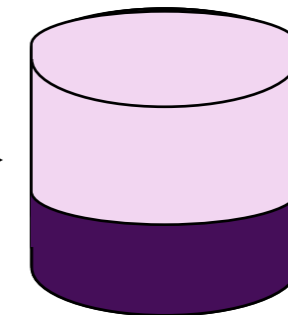
Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Registry

Portal



Index



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

Use case for TDWG standards - with index/cache

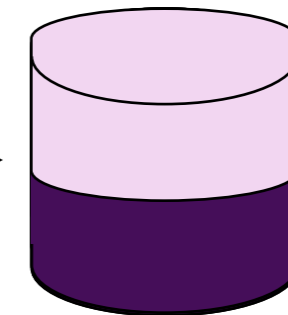
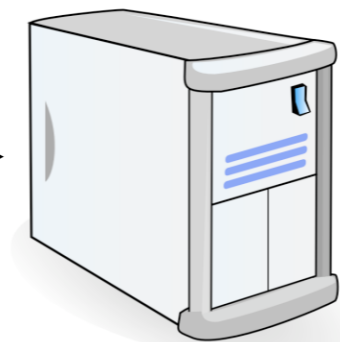


User A

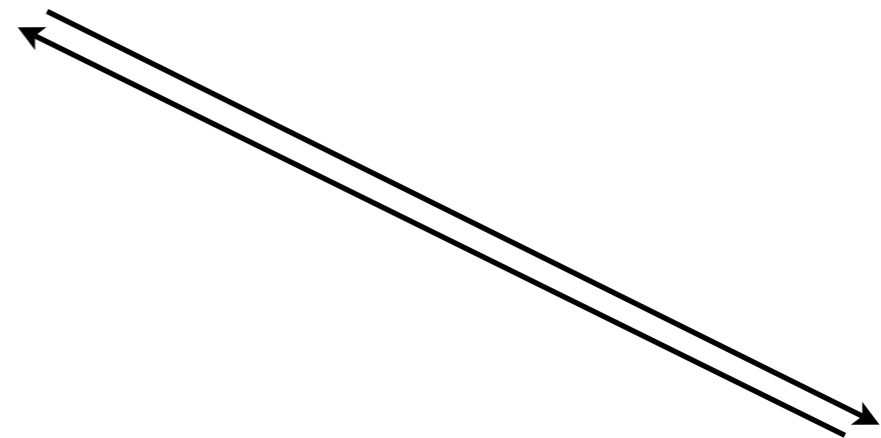
Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Registry

Portal



Index



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

Use case for TDWG standards - with index/cache

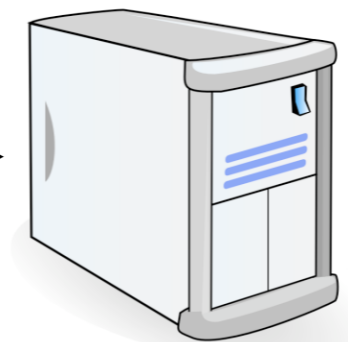


User A

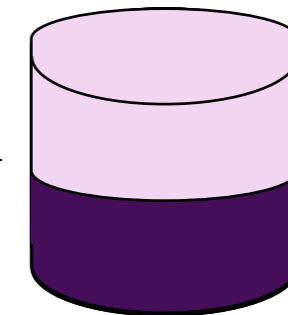
Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Registry

Portal



Index



Mammal Collection A



Mammal Collection B



Zoological Museum



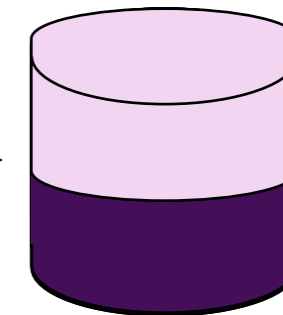
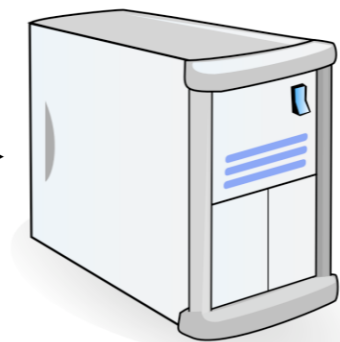
Botanic Garden

Use case for TDWG standards - with index/cache



User A

Request: Find Chiroptera
occurrences prior to 1950,
only using index



Registry

Portal

Index

Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

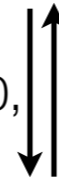
Use case for TDWG standards - with index/cache



User A

Request: Find Chiroptera occurrences prior to 1950, only using index

Response: 19 of 19 index records

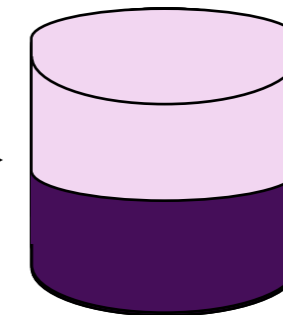
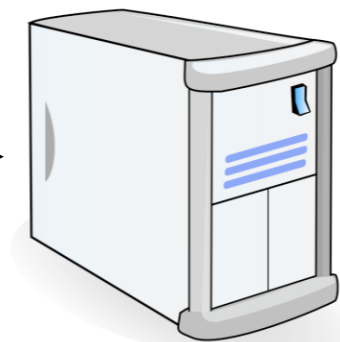


Registry

Portal

Index

Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

Use case for TDWG standards - with index/cache

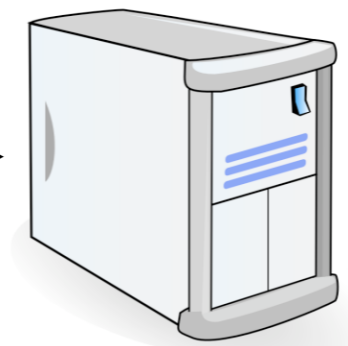


User A

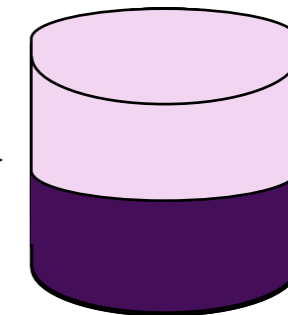
Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Registry

Portal



Index



Mammal Collection A



Mammal Collection B



Zoological Museum



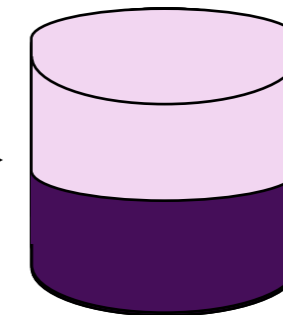
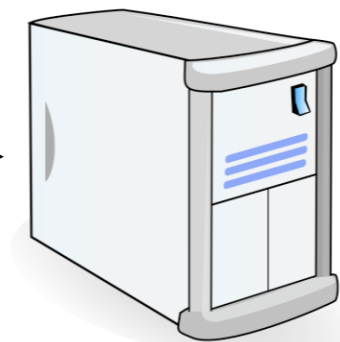
Botanic Garden

Use case for TDWG standards - with index/cache



User A

Request: Find Chiroptera occurrences prior to 1950, from original datasets



Registry

Portal

Index

Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae



Mammal Collection A



Mammal Collection B



Zoological Museum



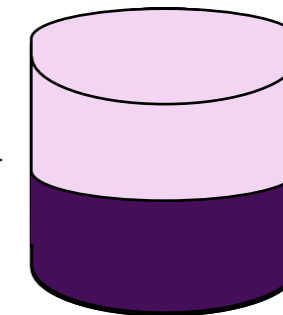
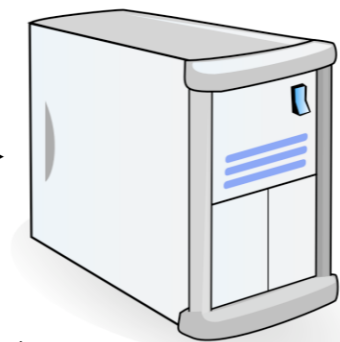
Botanic Garden

Use case for TDWG standards - with index/cache



User A

Request: Find Chiroptera occurrences prior to 1950, from original datasets



Registry

Portal

Index

Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae



Mammal Collection A



Mammal Collection B



Zoological Museum



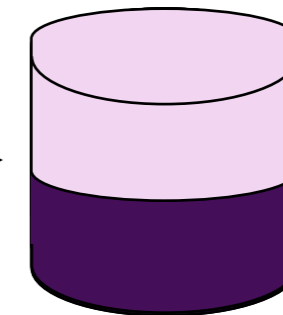
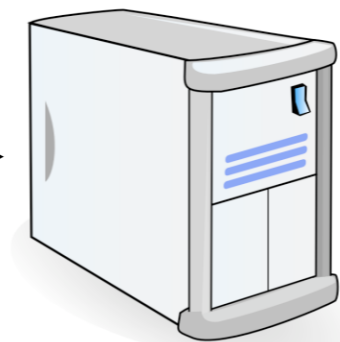
Botanic Garden

Use case for TDWG standards - with index/cache



User A

Request: Find Chiroptera occurrences prior to 1950, from original datasets



Registry

Portal

Index

Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae



Mammal Collection A



Mammal Collection B

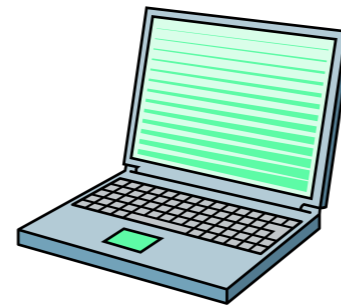


Zoological Museum



Botanic Garden

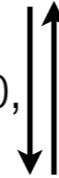
Use case for TDWG standards - with index/cache



User A

Request: Find Chiroptera occurrences prior to 1950, from original datases

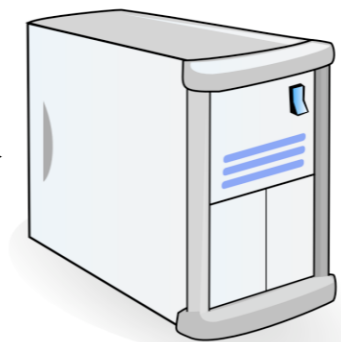
Response: 19 of 19 records



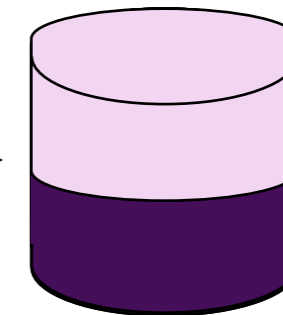
Key	Dataset	Protocol	Scope
1	Mammal Collection A	DiGIR	Mammalia
2	Mammal Collection B	DiGIR	Mammalia
3	Zoological Museum	DiGIR	Animalia
4	Botanic Garden	DiGIR	Plantae

Registry

Portal



Index



Mammal Collection A



Mammal Collection B



Zoological Museum



Botanic Garden

Key use case for Species Profile Model

- The decision whether a network should use an index/cache will depend on several factors:
 - The size of the network
 - As the network grows, indexing becomes more important*
 - The robustness and availability of the data providers
 - If some nodes may be offline, caching becomes more important*
 - Whether the data providers have the desire/capabilities to maintain a live server
 - Some providers may be happy just to upload data to a central store*
 - Whether search requests require joins between multiple data sets
 - Compound searches may need data to be located in one place*
 - Whether pre-processing data will make queries more reliable
 - Searches may be more complete against standardised data*

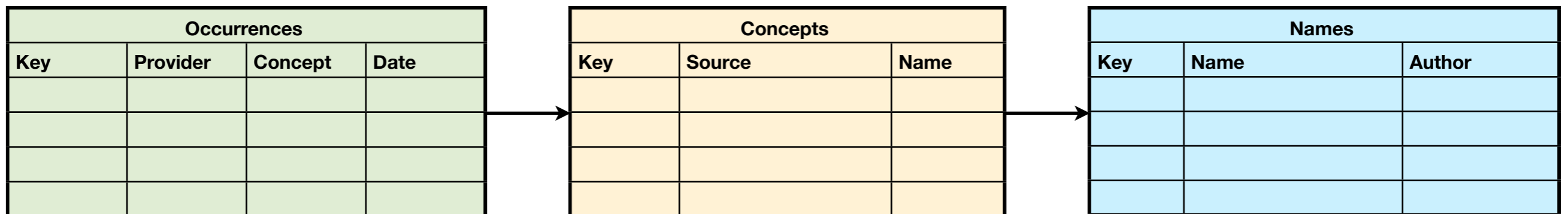
Implications for TDWG from index/cache model

- Need recommendations on the use of TAPIR and/or OAI-PMH for maintaining central caches of records
 - *Minimise the burden on data providers*
 - *Simplify task for indexing*
- Minimise alternative representations for same elements and stabilise key concepts
 - *Building a usable index depends on standardised data*
- Adopt or develop metadata standards for on-line datasets
 - *Taxonomic/geographic/temporal coverage and methods*
 - *References to standard taxonomies and vocabularies*
 - *Investigate EML as a model or standard for adoption*
- Make it easy to support attribution for each individual record
 - *Request/response models must allow different records to be related to different sets of metadata*

Integration using the TDWG ontology

Crawling data from web

Building a cache



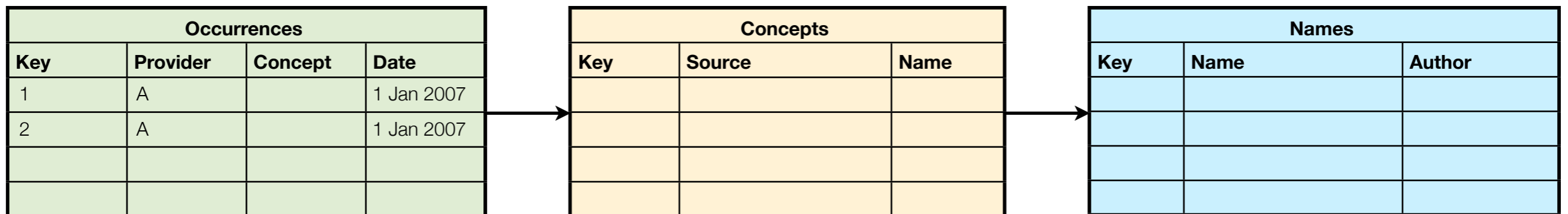
Integration using the TDWG ontology

Crawling data from web

Provider A

TAPIR Response - Occurrence Records	
GUID	urn:lsid:an.org:or:A01
ScientificName	<i>Quercus ilex</i>
DateCollected	1 Jan 2007
IdentifiedAs	urn:lsid:col.org:tc:123
GUID	urn:lsid:an.org:or:A01
ScientificName	<i>Pinus sylvestris</i>
DateCollected	1 Jan 2007
IdentifiedAs	urn:lsid:col.org:tc:456

Building a cache



Integration using the TDWG ontology

Crawling data from web

Provider A

Provider B

TAPIR Response - Occurrence Records

GUID	urn:lsid:an.org:or:A01
ScientificName	<i>Quercus ilex</i>
DateCollected	1 Jan 2007
IdentifiedAs	urn:lsid:col.org:tc:123

GUID	urn:lsid:an.org:or:A01
ScientificName	<i>Pinus sylvestris</i>
DateCollected	1 Jan 2007
IdentifiedAs	urn:lsid:col.org:tc:456

LSID Metadata - Taxon Concept

GUID	urn:lsid:col.org:tc:123
ScientificName	<i>Quercus ilex</i>
TaxonName	urn:lsid:ipni.org:tn:101
ParentConcept	urn:lsid:col.org:tc:120

Building a cache

Occurrences			
Key	Provider	Concept	Date
1	A	1	1 Jan 2007
2	A	2	1 Jan 2007

Concepts		
Key	Source	Name
1	CoL	1
2	CoL	2

Names		
Key	Name	Author
1	<i>Quercus ilex</i>	
2	<i>Pinus sylvestris</i>	

Integration using the TDWG ontology

Crawling data from web

Provider A

TAPIR Response - Occurrence Records

GUID	urn:lsid:an.org:or:A01
ScientificName	<i>Quercus ilex</i>
DateCollected	1 Jan 2007
IdentifiedAs	urn:lsid:col.org:tc:123

GUID	urn:lsid:an.org:or:A01
ScientificName	<i>Pinus sylvestris</i>
DateCollected	1 Jan 2007
IdentifiedAs	urn:lsid:col.org:tc:456

Provider B

LSID Metadata - Taxon Concept

GUID	urn:lsid:col.org:tc:123
ScientificName	<i>Quercus ilex</i>
TaxonName	urn:lsid:ipni.org:tn:101
ParentConcept	urn:lsid:col.org:tc:120

LSID Metadata - Taxon Concept

GUID	urn:lsid:col.org:tc:120
ScientificName	<i>Quercus</i>
ParentConcept	urn:lsid:col.org:tc:118

Building a cache

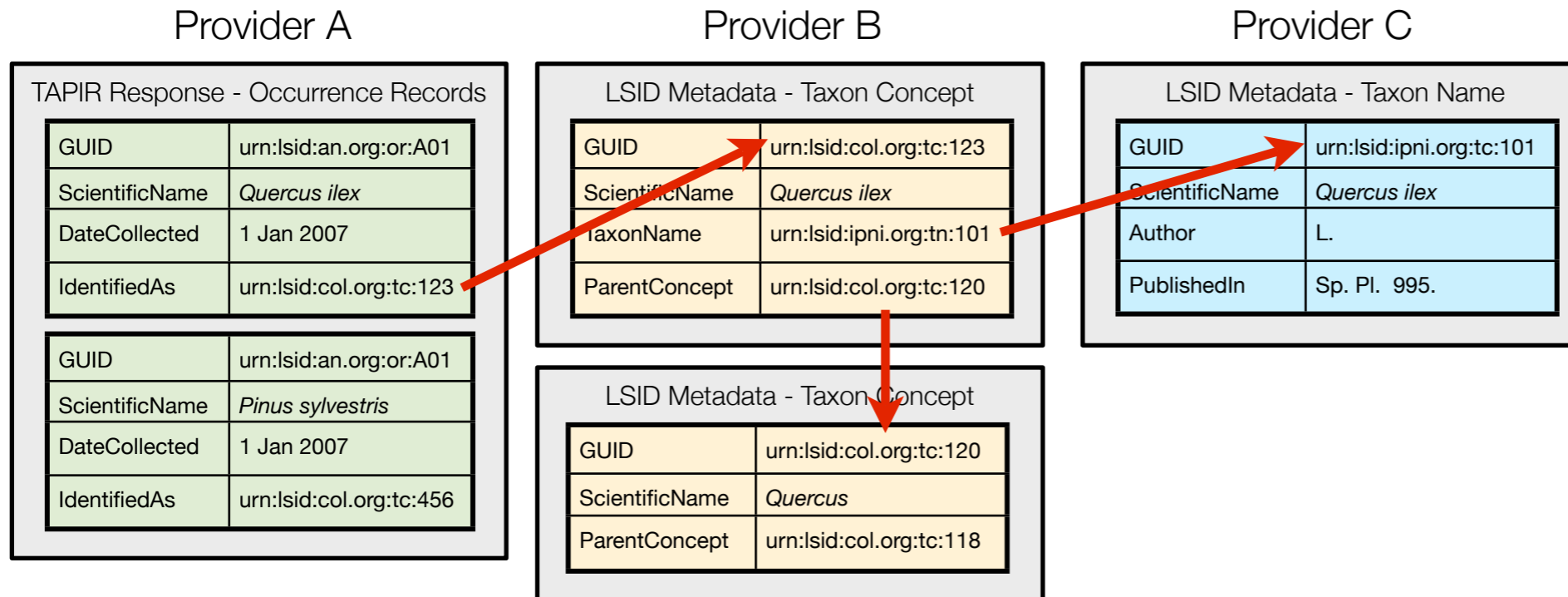
Occurrences			
Key	Provider	Concept	Date
1	A	1	1 Jan 2007
2	A	2	1 Jan 2007

Concepts		
Key	Source	Name
1	CoL	1
2	CoL	2
3	CoL	3

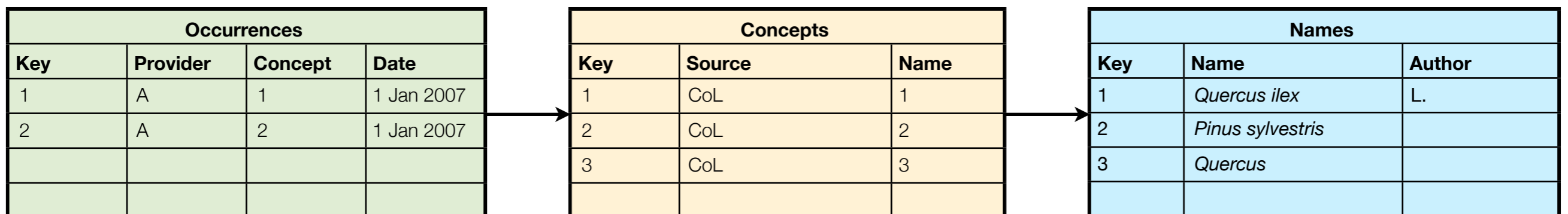
Names		
Key	Name	Author
1	<i>Quercus ilex</i>	
2	<i>Pinus sylvestris</i>	
3	<i>Quercus</i>	

Integration using the TDWG ontology

Crawling data from web

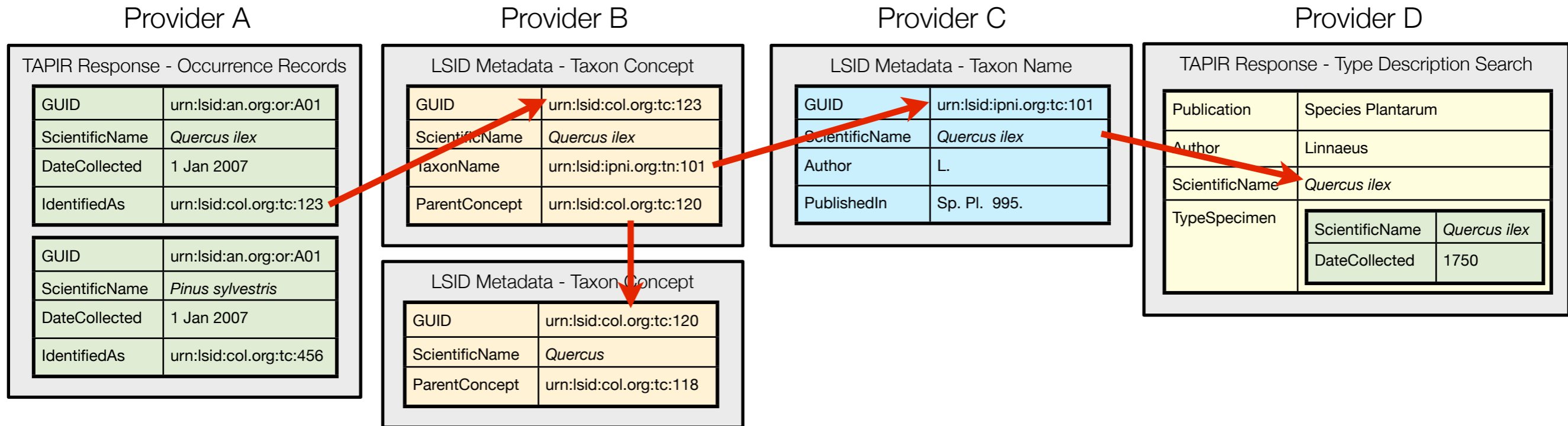


Building a cache

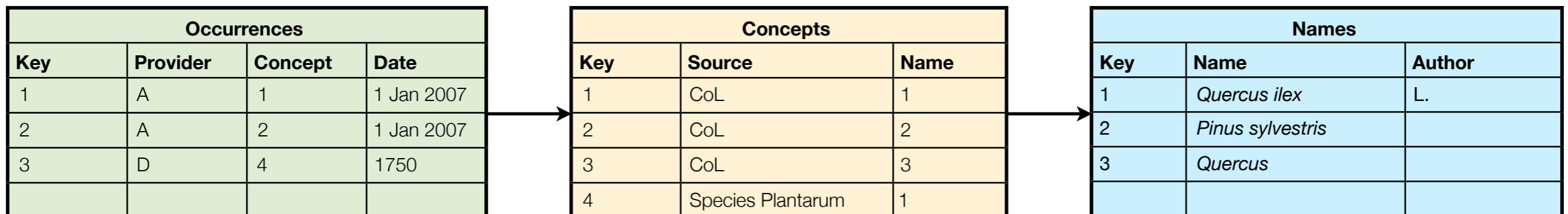


Integration using the TDWG ontology

Crawling data from web



Building a cache





GLOBAL
BIODIVERSITY
INFORMATION
FACILITY

Biodiversity
Information
Standards
T D W G

Thank you

Donald Hobern
Deputy Director for Informatics
GBIF Secretariat
Universitetsparken 15
2100 København Ø
Denmark

dhobern@gbif.org