GLOBAL
BIODIVERSITY
INFORMATION
FACILITY

Biodiversity
Information
Standards
TDWG

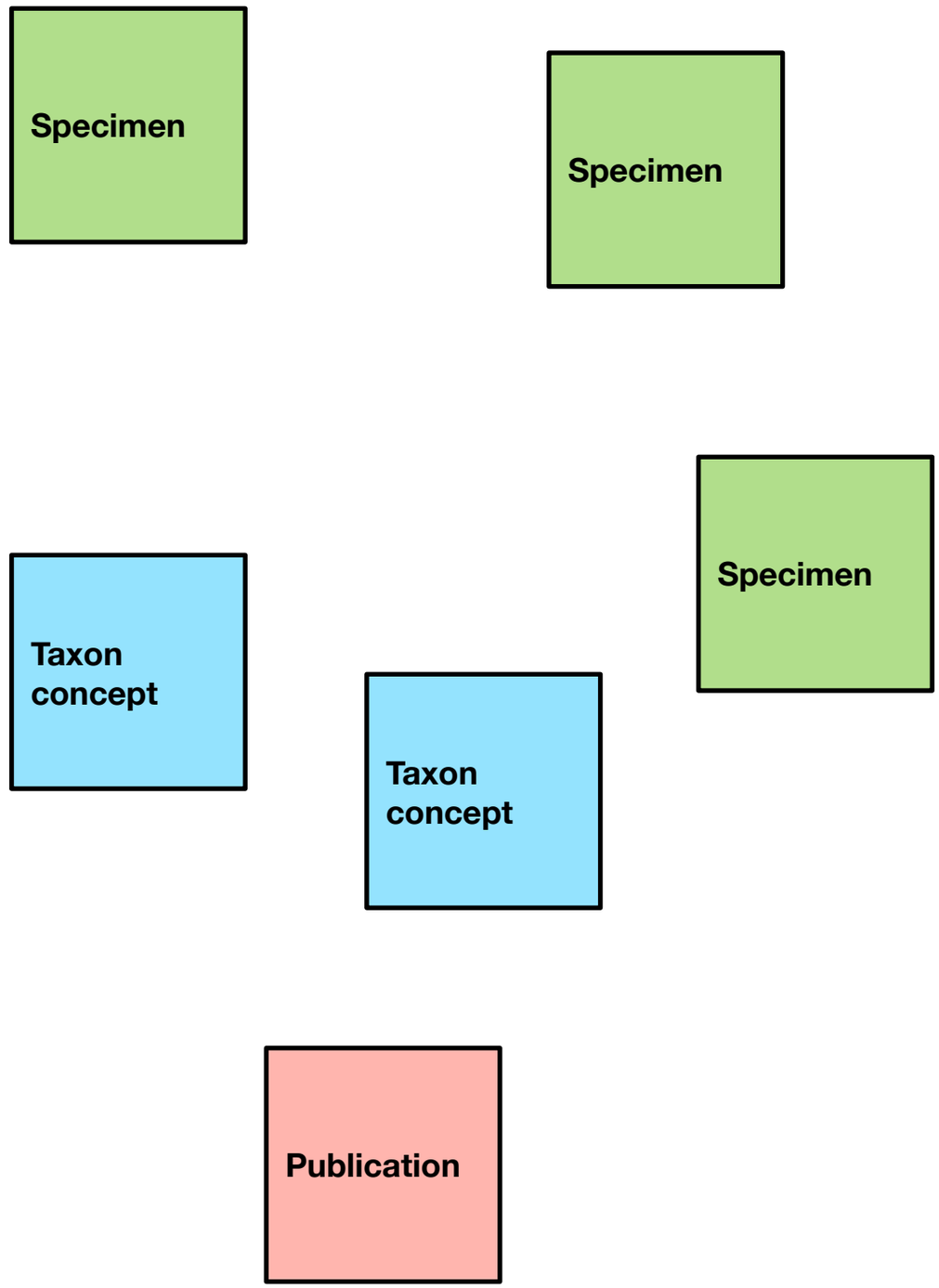
Key Enabling Technologies - Transfer Protocols

Donald Hobern - GBIF Deputy Director for Informatics
dhobern@gbif.org

September 2007

TDWG architecture - three core abilities

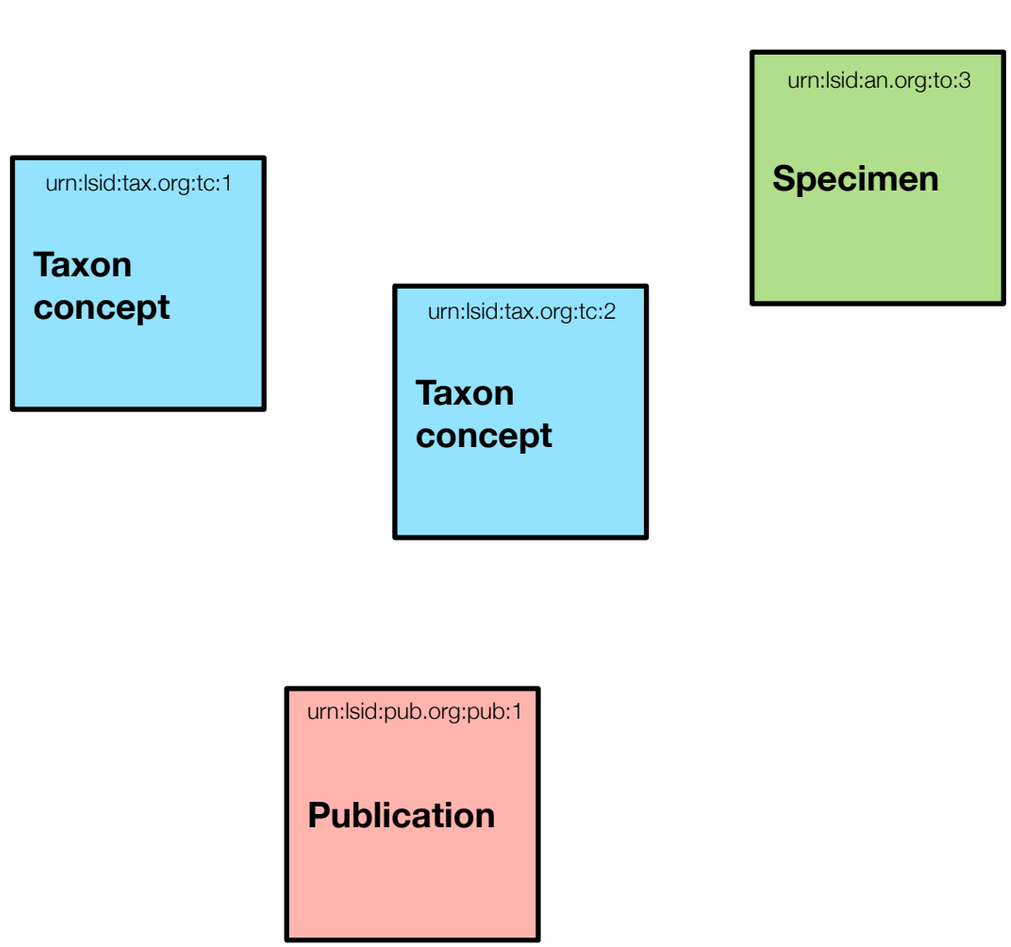
1. Construct data objects to model objects and concepts in biodiversity informatics
    - This is the role of the TDWG data standards
    - Standards being recast as reusable (LSID) vocabularies to simplify reuse



TDWG architecture - three core abilities

1. Construct data objects to model objects and concepts in biodiversity informatics
   - This is the role of the TDWG data standards
   - Standards being recast as reusable (LSID) vocabularies to simplify reuse

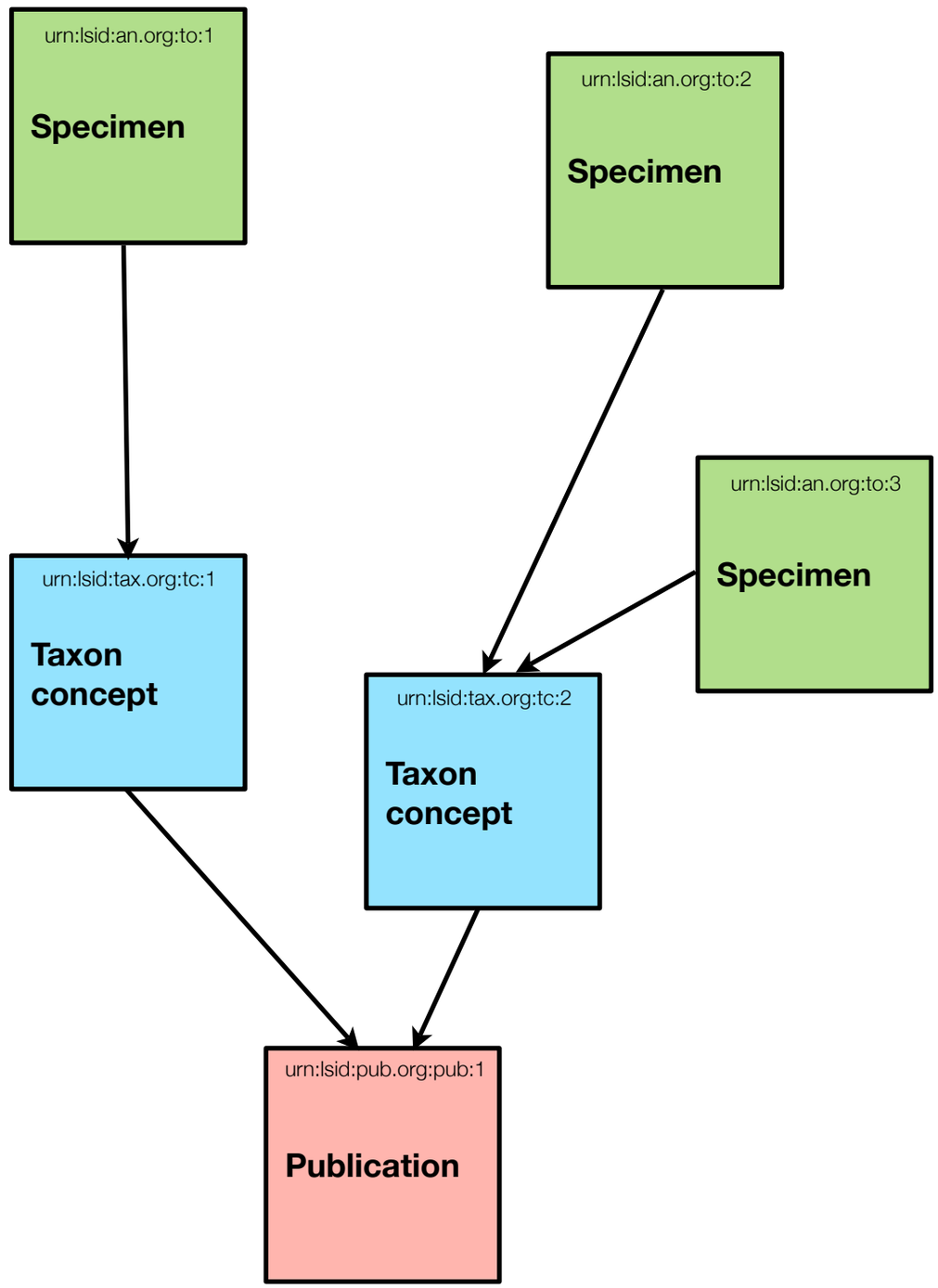
2. Refer reliably to data objects - make them linkable
   - This is the role of globally unique identifiers (GUIDs)
   - TDWG has adopted Life Science Identifiers (LSIDs) - compatible with the TDWG vocabularies



TDWG architecture - three core abilities

1. Construct data objects to model objects and concepts in biodiversity informatics
   - This is the role of the TDWG data standards
   - Standards being recast as reusable (LSID) vocabularies to simplify reuse

2. Refer reliably to data objects - make them linkable
   - This is the role of globally unique identifiers (GUIDs)
   - TDWG has adopted Life Science Identifiers (LSIDs) - compatible with the TDWG vocabularies

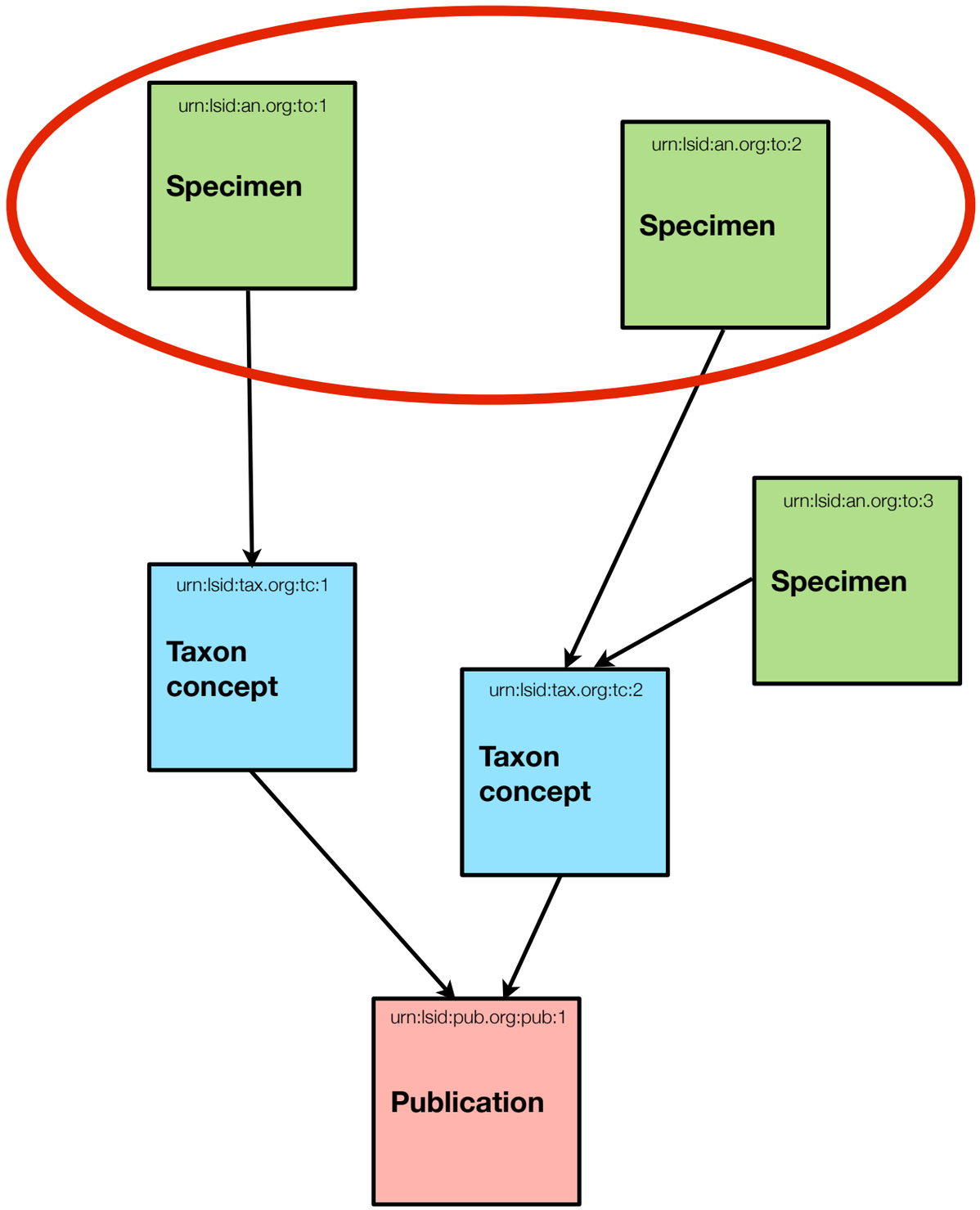


TDWG architecture - three core abilities

1. Construct data objects to model objects and concepts in biodiversity informatics
   - This is the role of the TDWG data standards
   - Standards being recast as reusable (LSID) vocabularies to simplify reuse

2. Refer reliably to data objects - make them linkable
   - This is the role of globally unique identifiers (GUIDs)
   - TDWG has adopted Life Science Identifiers (LSIDs) - compatible with the TDWG vocabularies

3. Discover and access data objects
   - This is the role of the access protocols
   - TDWG has developed DiGIR and BioCASe as access protocols and now developed TAPIR as their successor



TDWG data access protocols

- DiGIR
  - Most widely used TDWG protocol (including MaNIS, ORNIS, HerpNET, OBIS, GBIF)
  - Usually returning Darwin Core data (occurrence records with simple structure)
  - Implementations in PHP and Python (and some special adaptors for particular systems)
  - Metadata, inventory and search operations
- BioCASe
  - Large number of providers in Europe - the BioCASE project - and accessed by GBIF
  - Usually returning ABCD data (occurrence records with complex structure)
  - Implementation in Python
  - Capabilities, scan (== inventory) and search operations
- TAPIR
  - Early providers already connected to GBIF - protocol adopted for use by several networks
  - Returning Darwin Core or ABCD data (can support a range of structures)
  - Exploring use with new TDWG vocabularies
  - Implementations in Python, PHP and .NET
  - Metadata, capabilities, ping, inventory and search operations

TAPIR inventory

Example request

```
<?xml version="1.0" encoding="UTF-8"?>
<request xmlns="http://rs.tdwg.org/tapir/1.0"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://rs.tdwg.org/tapir/1.0
                             http://rs.tdwg.org/tapir/1.0/schema/tapir.xsd">
  <header>
    <source sendtime="2005-11-11T12:23:56.023+01:00">
    </source>
  </header>
  <inventory count="true" limit="100" start="0">
    <concepts>
      <concept id="http://example.net/schema1/Country" tagName="country"/>
      <concept id="http://example.net/schema1/Genus" tagName="genus"/>
    </concepts>
    <filter>
      <like>
        <concept id="http://example.net/schema1/Genus"/>
        <literal value="Luzu*"/>
      </like>
    </filter>
  </inventory>
</request>
```

TAPIR inventory

Example request

```
<?xml version="1.0" encoding="UTF-8"?>
<request xmlns="http://rs.tdwg.org/tapir/1.0"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://rs.tdwg.org/tapir/1.0
                             http://rs.tdwg.org/tapir/1.0/schema/tapir.xsd">
  <header>
    <source sendtime="2005-11-11T12:23:56.023+01:00">
    </source>
  </header>
  <inventory count="true" limit="100" start="0">
    <concepts>
      <concept id="http://example.net/schema1/Country" tagName="country"/>
      <concept id="http://example.net/schema1/Genus" tagName="genus"/>
    </concepts>
    <filter>
      <like>
        <concept id="http://example.net/schema1/Genus"/>
        <literal value="Luzu*"/>
      </like>
    </filter>
  </inventory>
</request>
```

Example URL-encoded request

```
http://example.net/tapir.cgi?op=inventory&count=false&start=0&limit=100& concept=Country@schema1&concept=Genus@schema1
```

TAPIR inventory

Example request

```
<?xml version="1.0" encoding="UTF-8"?>
<request xmlns="http://rs.tdwg.org/tapir/1.0"
         xmlns:xsi="http://www.w3.org/2001/XM
         xsi:schemaLocation="http://rs.tdwg.o
                            http://rs.tdwg.o
  <header>
    <source sendtime="2005-11-11T12:23:56.023
    </source>
  </header>
  <inventory count="true" limit="100" start="
    <concepts>
      <concept id="http://example.net/schema1
      <concept id="http://example.net/schema1
    </concepts>
    <filter>
      <like>
        <concept id="http://example.net/schem
        <literal value="Luzu*"/>
      </like>
    </filter>
  </inventory>
</request>
```

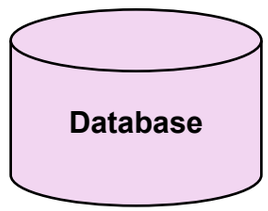
Example response

```
<?xml version="1.0" encoding="UTF-8"?>
<response xmlns="http://rs.tdwg.org/tapir/1.0"
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:schemaLocation="http://rs.tdwg.org/tapir/1.0
                              http://rs.tdwg.org/tapir/1.0/schema/
tapir.xsd">
  <header>
    <source accesspoint="http://example.net/tapir.cgi"
            sendtime="2005-11-11T12:23:56.023+01:00">
      <software name="TapirService" version="1.0"/>
    </source>
  </header>
  <inventory>
    <concepts>
      <concept id="http://example.net/schema1/Country"/>
      <concept id="http://example.net/schema1/Genus"/>
    </concepts>
    <record>
      <value>AUSTRALIA</value>
      <value>Calicium</value>
    </record>
    <record>
      <value>AUSTRALIA</value>
      <value>Fellhanera</value>
    </record>
    <summary start="0" next="2" totalReturned="2" totalMatched="35"/>
  </inventory>
</response>
```

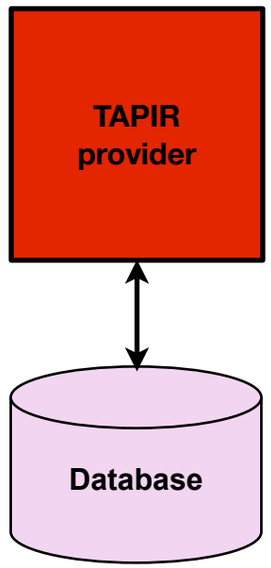
Example URL-encoded request

http://example.net/tapir.cgi?op=inventory&count=false&start=0&limit=100& concept=Country@schema1&concept=Genus@schema1

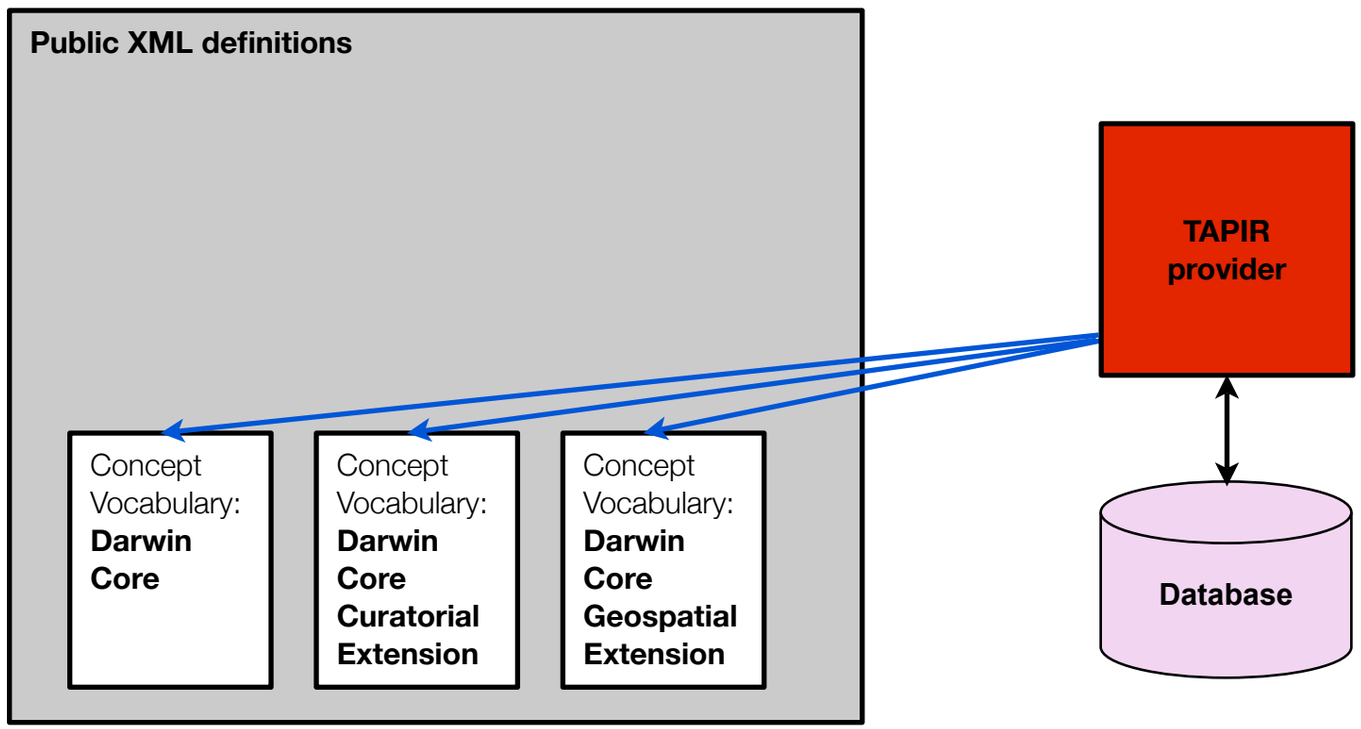
TAPIR - a flexible interface to data



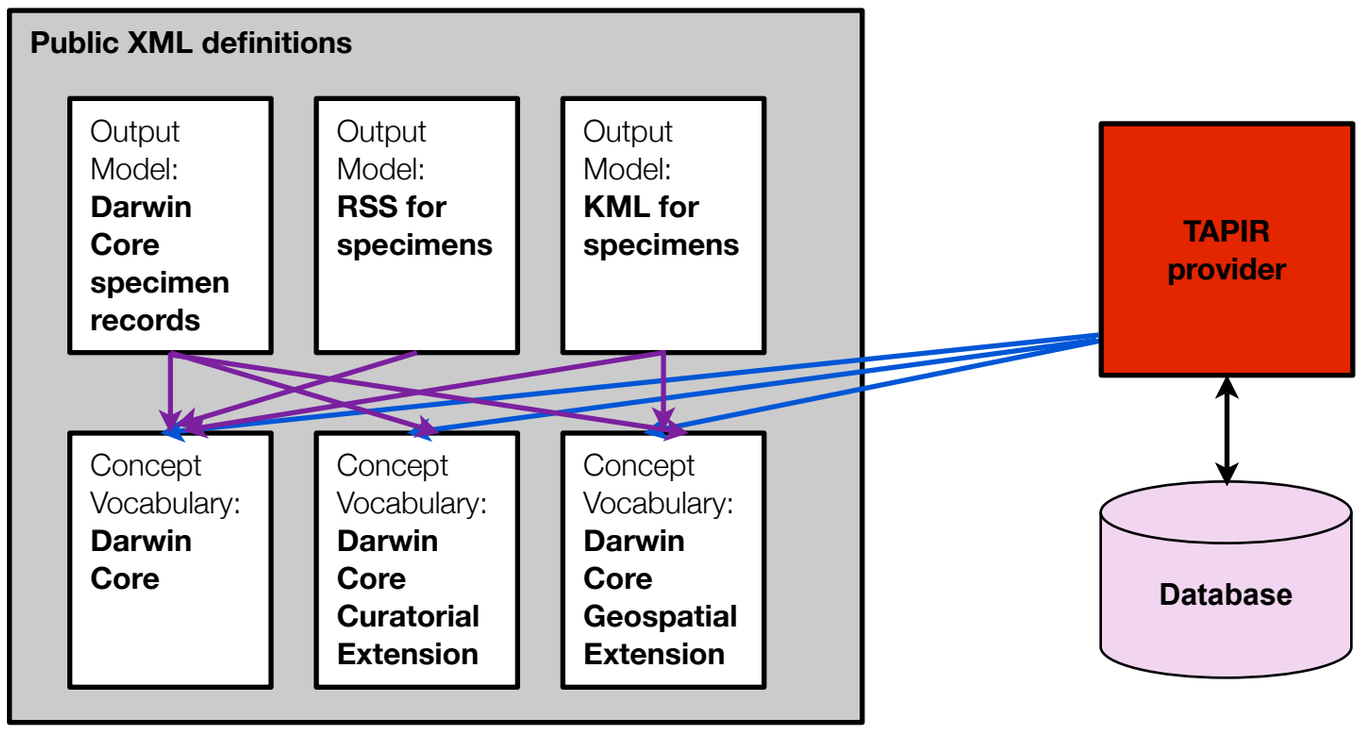
TAPIR - a flexible interface to data



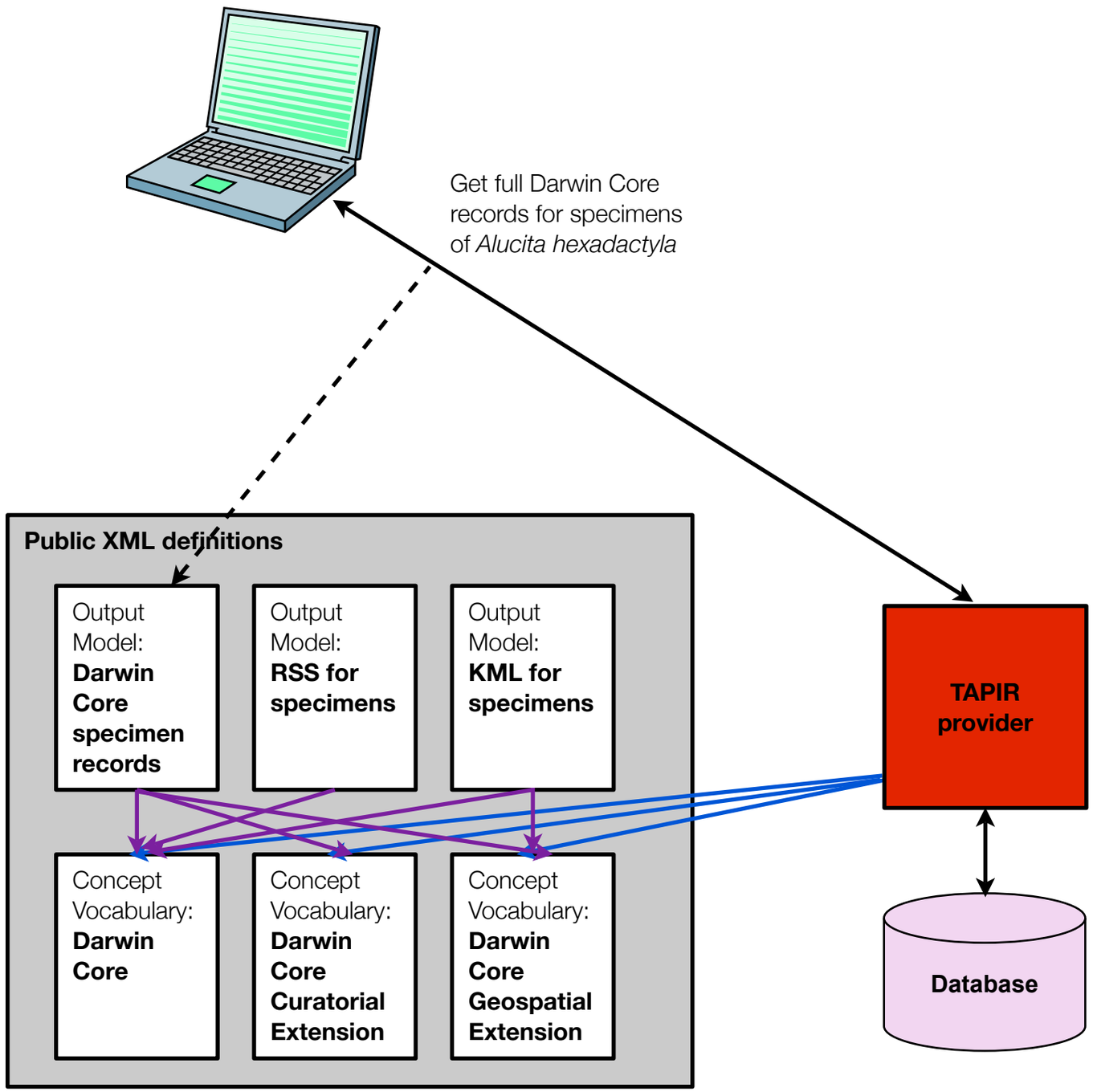
TAPIR - a flexible interface to data



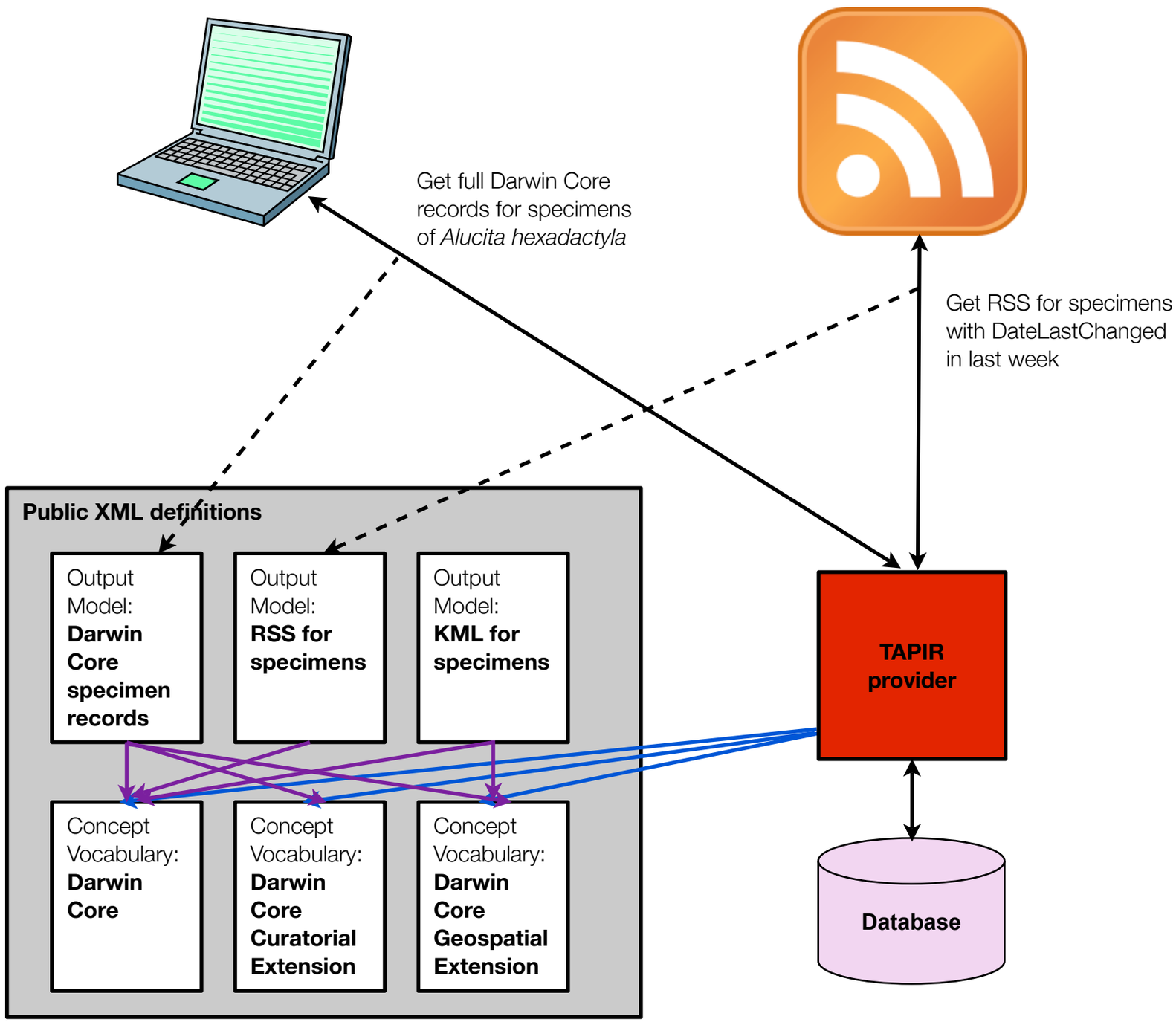
TAPIR - a flexible interface to data



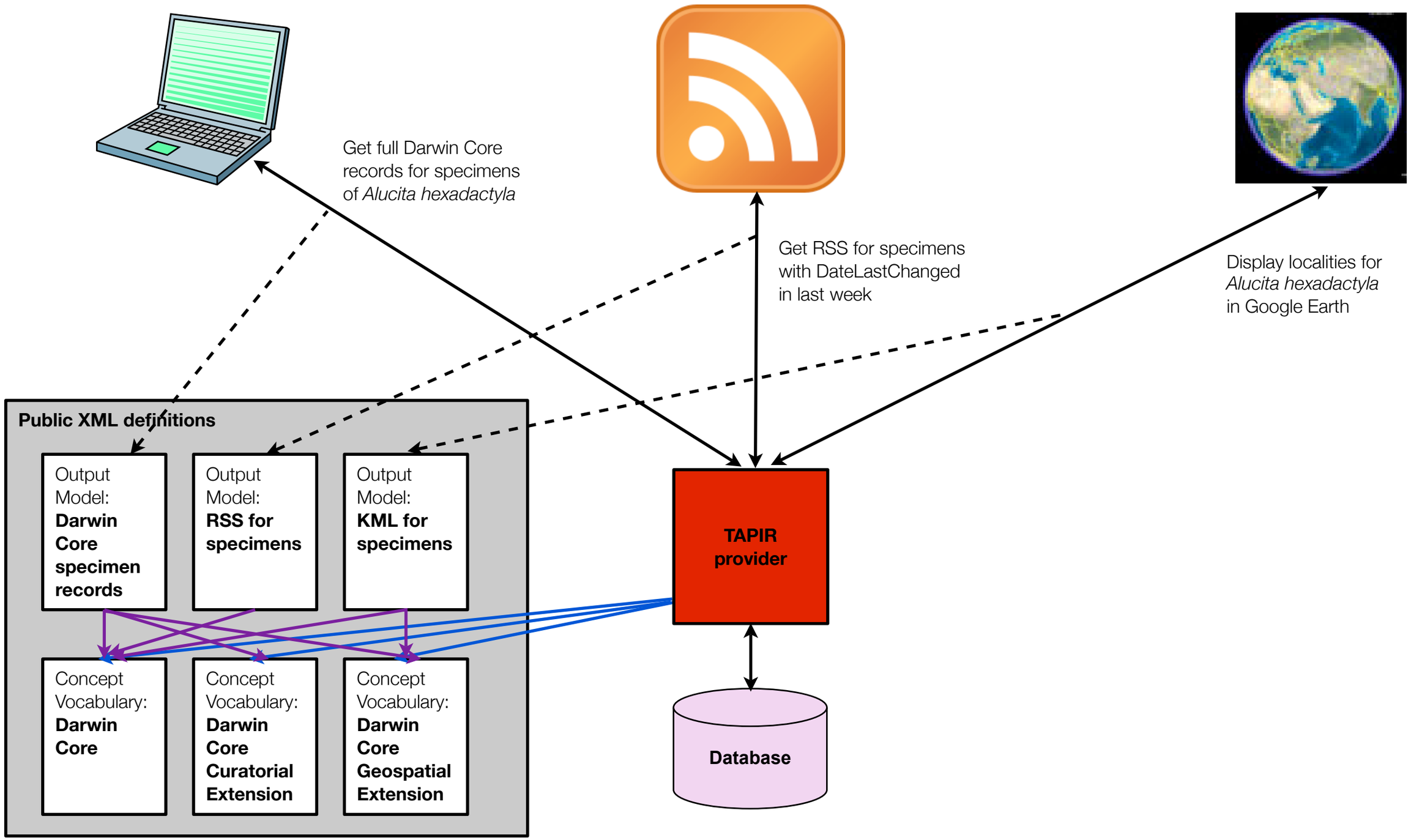
TAPIR - a flexible interface to data



TAPIR - a flexible interface to data



TAPIR - a flexible interface to data



Other protocols: OAI-PMH

- Simple protocol for harvesting metadata from a data provider
- A data provider serves a **repository** of **record**s each with an **identifier** and makes the records available in various **metadata formats**
- Six request types
  - **Identify** - Get information describing a repository
  - **ListMetadataFormats** - List supported metadata formats
  - **ListIdentifiers** - List identifiers of available records (e.g. records changed since a given date)
  - **ListRecords** - List full records (e.g. records changed since a given date)
  - **GetRecord** - Get the record with a given identifier (in a given metadata format)
  - **ListSets** - List supported sets (categories of record - these can be used to limit results with ListIdentifiers and ListRecords)

OAI-PMH - a protocol for harvesting metadata

Identify

Repository

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
        http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2002-02-08T12:00:01Z</responseDate>
  <request verb="Identify">http://memory.loc.gov/cgi-bin/oai</request>
  <Identify>
    <repositoryName>Library of Congress Open Archive Initiative
                Repository 1</repositoryName>
    <baseURL>http://memory.loc.gov/cgi-bin/oai</baseURL>
    <adminEmail>anybody@loc.gov</adminEmail>
    <earliestDatestamp>1990-02-01T12:00:00Z</earliestDatestamp>
    ...
```

Identifier	Metadata	Date	Sets
urn:lsid:my.org:item:12646	<oai_dc:dc ...	14/09/07	
urn:lsid:my.org:item:12645	<oai_dc:dc ...	14/09/07	
urn:lsid:my.org:item:12644	<oai_dc:dc ...	14/09/07	
urn:lsid:my.org:item:12643	<oai_dc:dc ...	11/09/07	
urn:lsid:my.org:item:12642	<oai_dc:dc ...	11/09/07	
urn:lsid:my.org:item:12641	<oai_dc:dc ...	03/09/07	
urn:lsid:my.org:item:12640	<oai_dc:dc ...	02/09/07	
urn:lsid:my.org:item:12639	<oai_dc:dc ...	25/08/07	
urn:lsid:my.org:item:12638	<oai_dc:dc ...	25/08/07	
urn:lsid:my.org:item:12639	<oai_dc:dc ...	25/08/07	

OAI-PMH - a protocol for harvesting metadata

ListIdentifiers
from 7 Sep 2007

Repository

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
        http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
 <responseDate>2002-02-08T12:00:01Z</responseDate>
 <request verb="Identify">http://memory.loc.gov/cgi-bin/oai</request>
 <Identify>
  <repositoryName>Library of Congress Open Archive Initiative
            Repository 1</repositoryName>
  <baseURL>http://memory.loc.gov/cgi-bin/oai</baseURL>
  <adminEmail>anybody@loc.gov</adminEmail>
  <earliestDatestamp>1990-02-01T12:00:00Z</earliestDatestamp>
  ...
```

Identifier	Metadata	Date	Sets
urn:lsid:my.org:item:12646	<oai_dc:dc ...	14/09/07	
urn:lsid:my.org:item:12645	<oai_dc:dc ...	14/09/07	
urn:lsid:my.org:item:12644	<oai_dc:dc ...	14/09/07	
urn:lsid:my.org:item:12643	<oai_dc:dc ...	11/09/07	
urn:lsid:my.org:item:12642	<oai_dc:dc ...	11/09/07	
urn:lsid:my.org:item:12641	<oai_dc:dc ...	03/09/07	
urn:lsid:my.org:item:12640	<oai_dc:dc ...	02/09/07	
urn:lsid:my.org:item:12639	<oai_dc:dc ...	25/08/07	
urn:lsid:my.org:item:12638	<oai_dc:dc ...	25/08/07	
urn:lsid:my.org:item:12639	<oai_dc:dc ...	25/08/07	

OAI-PMH - a protocol for harvesting metadata

GetRecord

urn:lsid:my.org.item:12644

Repository

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
        http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
 <responseDate>2002-02-08T12:00:01Z</responseDate>
 <request verb="Identify">http://memory.loc.gov/cgi-bin/oai</request>
 <Identify>
  <repositoryName>Library of Congress Open Archive Initiative
            Repository 1</repositoryName>
  <baseURL>http://memory.loc.gov/cgi-bin/oai</baseURL>
  <adminEmail>anybody@loc.gov</adminEmail>
  <earliestDatestamp>1990-02-01T12:00:00Z</earliestDatestamp>
  ...
```

Identifier	Metadata	Date	Sets
urn:lsid:my.org:item:12646	<oai_dc:dc ...	14/09/07	
urn:lsid:my.org:item:12645	<oai_dc:dc ...	14/09/07	
urn:lsid:my.org:item:12644	<oai_dc:dc ...	14/09/07	
urn:lsid:my.org:item:12643	<oai_dc:dc ...	11/09/07	
urn:lsid:my.org:item:12642	<oai_dc:dc ...	11/09/07	
urn:lsid:my.org:item:12641	<oai_dc:dc ...	03/09/07	
urn:lsid:my.org:item:12640	<oai_dc:dc ...	02/09/07	
urn:lsid:my.org:item:12639	<oai_dc:dc ...	25/08/07	
urn:lsid:my.org:item:12638	<oai_dc:dc ...	25/08/07	
urn:lsid:my.org:item:12639	<oai_dc:dc ...	25/08/07	

OAI-PMH - a protocol for harvesting metadata

ListRecords
set=**GREEN**

Repository

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
        http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2002-02-08T12:00:01Z</responseDate>
  <request verb="Identify">http://memory.loc.gov/cgi-bin/oai</request>
  <Identify>
    <repositoryName>Library of Congress Open Archive Initiative
                Repository 1</repositoryName>
    <baseURL>http://memory.loc.gov/cgi-bin/oai</baseURL>
    <adminEmail>anybody@loc.gov</adminEmail>
    <earliestDatestamp>1990-02-01T12:00:00Z</earliestDatestamp>
    ...
```

Identifier	Metadata	Date	Sets
urn:lsid:my.org:item:12646	<oai_dc:dc ...	14/09/07	
urn:lsid:my.org:item:12645	<oai_dc:dc ...	14/09/07	
urn:lsid:my.org:item:12644	<oai_dc:dc ...	14/09/07	
urn:lsid:my.org:item:12643	<oai_dc:dc ...	11/09/07	
urn:lsid:my.org:item:12642	<oai_dc:dc ...	11/09/07	
urn:lsid:my.org:item:12641	<oai_dc:dc ...	03/09/07	
urn:lsid:my.org:item:12640	<oai_dc:dc ...	02/09/07	
urn:lsid:my.org:item:12639	<oai_dc:dc ...	25/08/07	
urn:lsid:my.org:item:12638	<oai_dc:dc ...	25/08/07	
urn:lsid:my.org:item:12639	<oai_dc:dc ...	25/08/07	

Other protocols: WFS and SPARQL

- Open Geospatial Consortium (OGC) Web Feature Service (WFS)
  - One of several OGC web service models
  - WFS implementations return geographical **features**
  - A **feature** is a data object which can be represented or manipulated by GIS tools
  - TDWG occurrence data can be represented as **features** using e.g. Darwin Core terms
  - A feature service implementation can support filtered queries like TAPIR
- SPARQL (SPARQL Protocol and RDF Query Language)
  - Supports queries (potentially across multiple distributed databases) based on RDF triple patterns
  - Potential example: *Find* ***TaxonOccurrence*** *objects* ***recordedAt ?locality*** *where* ***?locality isA ProtectedArea***
  - Very powerful but not suited for general use with our data at this stage

Why use different protocols?

- TAPIR
  - Powerful general search capability (complex filters)
  - Flexible representation of output data
  - Easy to handle new classes of data object
- OAI-PMH
  - Highly efficient for maintaining caches of data
  - Scope for supporting broad classes of user via set mechanism
- WFS
  - Essential for exposing data objects to GIS tools in a standards-compliant form
- SPARQL
  - Valuable for exploring repositories of RDF data and reasoning over their contents

GLOBAL
BIODIVERSITY
INFORMATION
FACILITY

Biodiversity
Information
Standards
TDWG

Thank you

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

dhobern@gbif.org