

Data Standards and Protocols for Biological Collections Data

Renato De Giovanni

CRIA – Reference Center on Environmental Information

renato@cria.org.br

ICCC12, September 2010

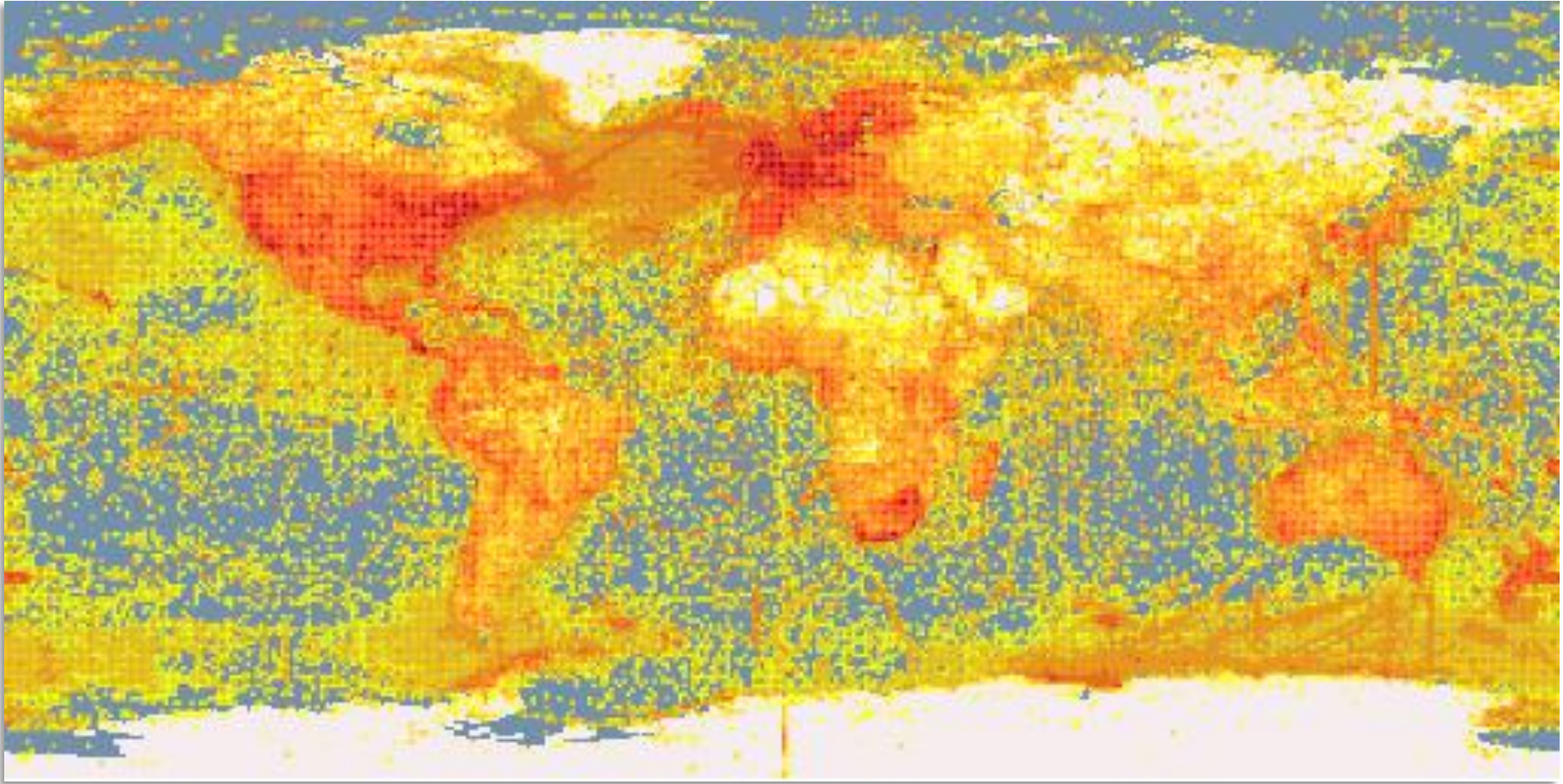
Outline

1. Introduction
2. Strategies for data integration
3. Data standards for biological collections
4. Query protocols
5. Final recommendations

Access to biodiversity data

- During the last 15 years, large quantities of primary biodiversity data became available:
 - Advances in informatics (large-capacity storage media, Internet, communication infrastructure).
 - Large-scale data digitization programmes for biological collections.
 - Initiatives to build data sharing networks.
- Number of networks keeps growing.

GBIF example



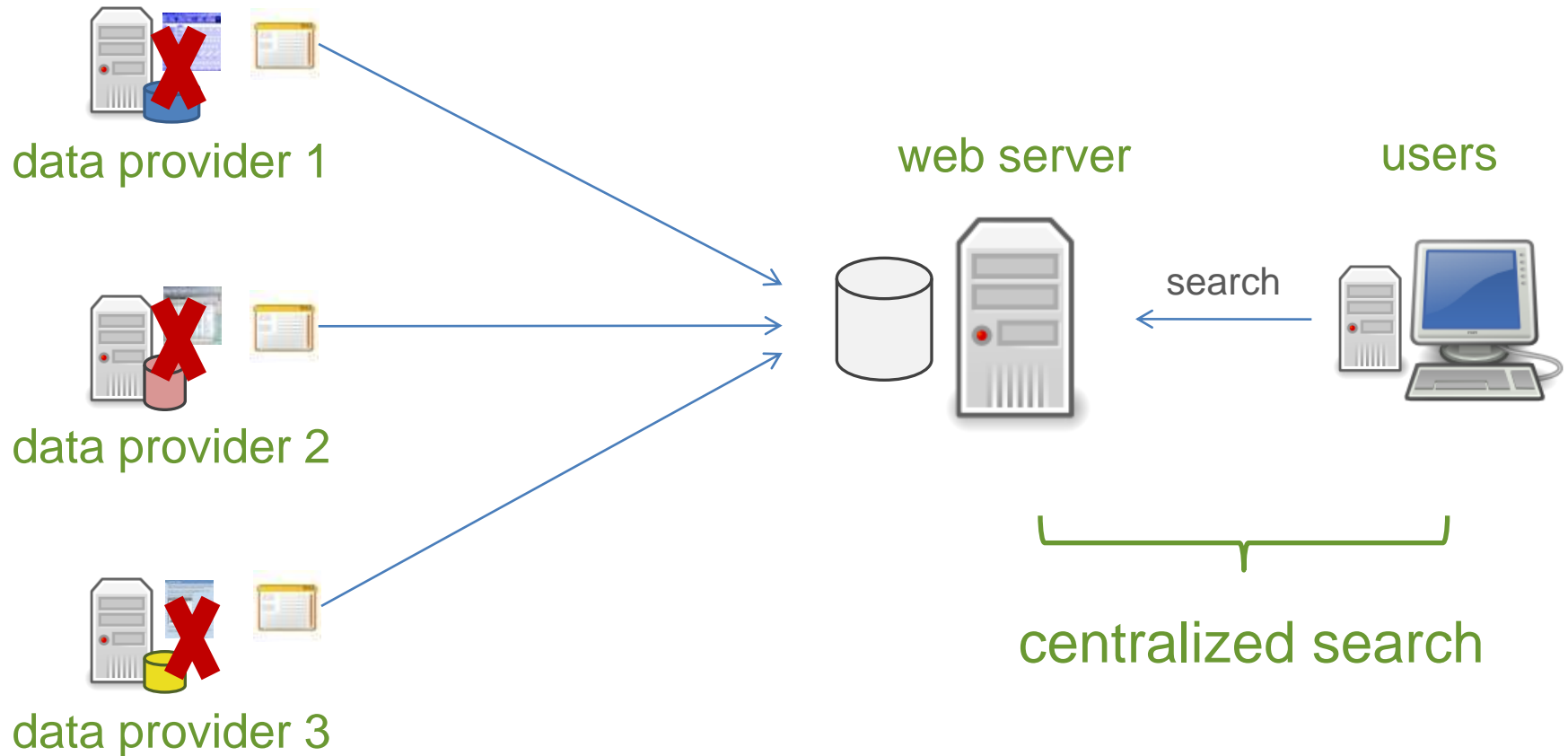
GBIF network: > 200M records

Building biodiversity data networks

- Technical, financial, social and political issues involved.
- Among the technical issues:
 - Definition of network architecture.
 - Adoption of data standards & protocols.

Strategies for data integration

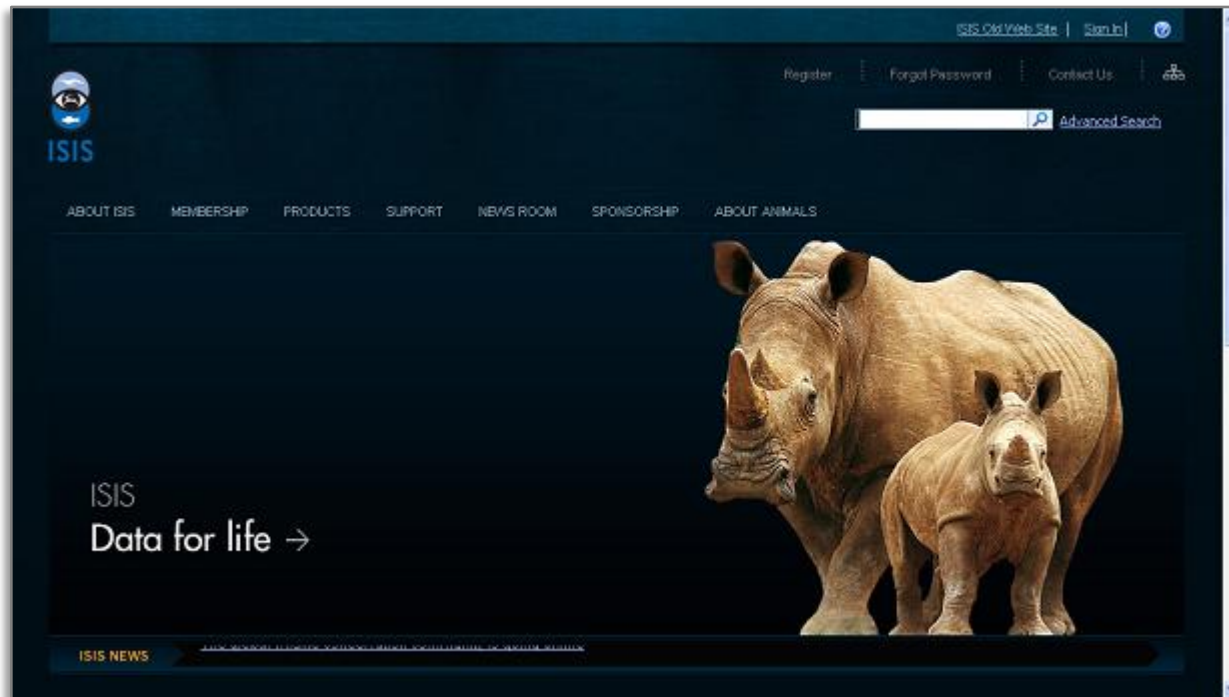
1) Same software and database used by all providers



Strategies for data integration

1) Same software and database used by all providers

Example: International Species Information System (ISIS)



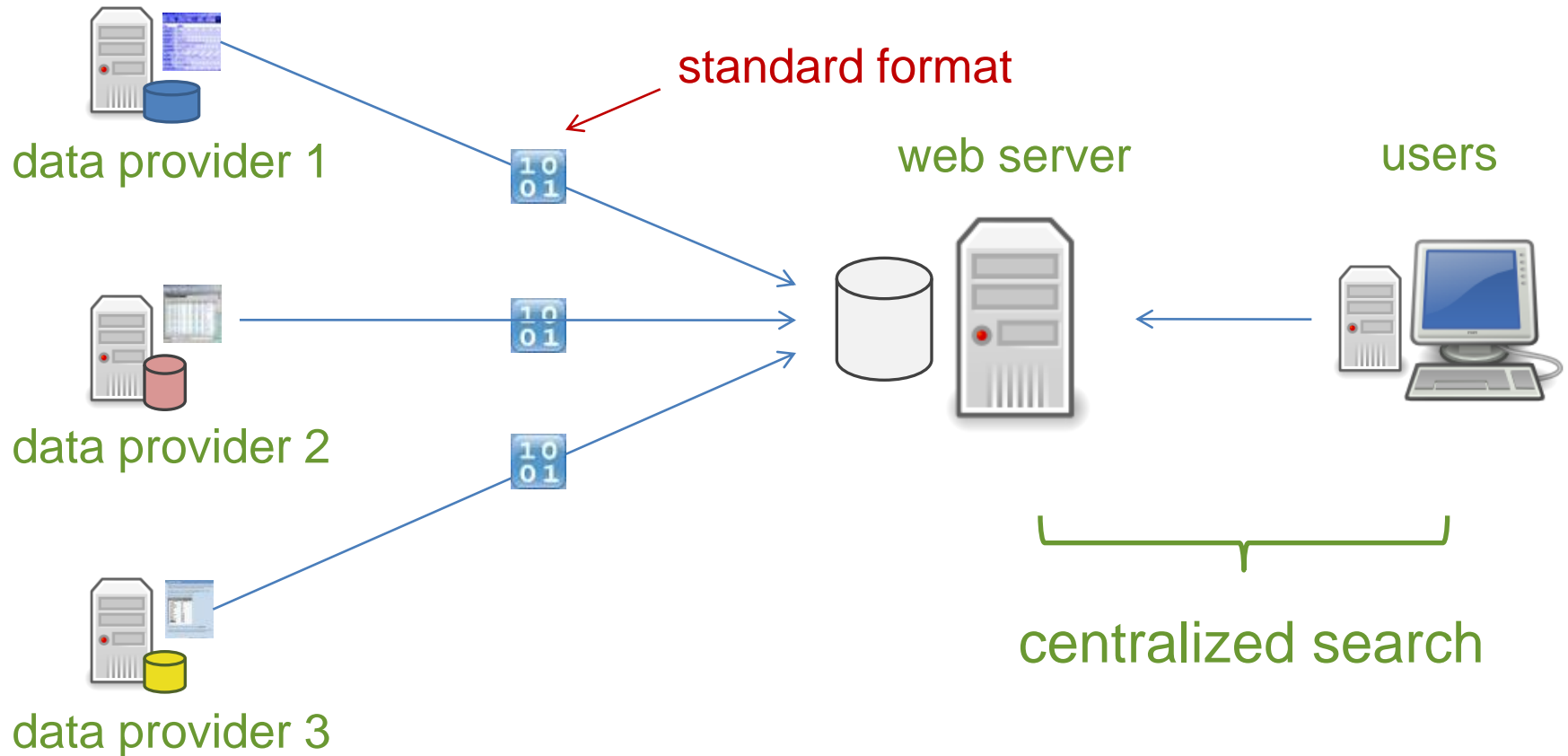
- 825 institutions (zoos and aquariums) from 76 countries.
- Initiative started in 1973.

Strategies for data integration

- 1) Same software and database used by all providers
 - ✓ Interesting solution if all providers agree to use the same system:
 - ✓ Improvements benefit all participants.
 - ✓ Shared costs.
 - ✓ Good performance (although queries are run in the production database).
 - Lack of freedom to make custom adjustments.
 - Very difficult to accomplish if providers are already using their own management software (sometimes developed with considerable effort).

Strategies for data integration

2) Periodically export data to a central database



Strategies for data integration

2) Periodically export data to a central database

Examples:



- **Common Access to Biological Resources and Information.**
- Started in 1999.
- 28 catalogues from European institutions (>100K records).



- ~60 BRCs.
- Includes screen scraping.



- Brazilian network.
- Recently switched to the next architecture...

Strategies for data integration

2) Periodically export data to a central database

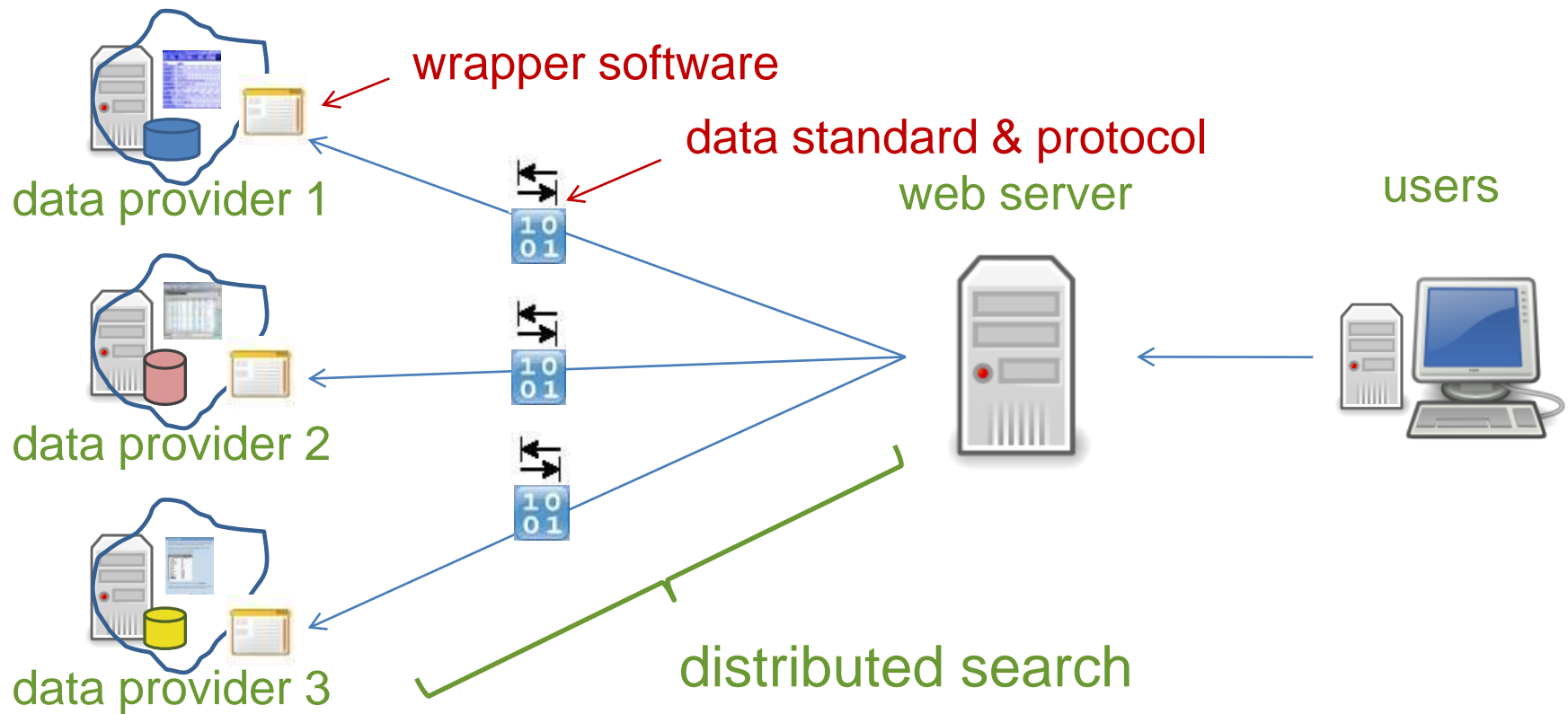
- ✓ Good performance.
- ✓ Easier to implement.
- Queries are performed on potentially non current data.
- Onus on providers to transform data into a common format and periodically export it.
- Risk of (quite) infrequent updates (SICol). However...



~7 million books from 1700 book stores!

Strategies for data integration

3) Real time distributed queries



Strategies for data integration

3) Real time distributed queries

Examples:



- 1998 - 2003.
- North America.
- MaNIS, HerpNET, ORNIS & FishNet.



REMIB

Red Mundial de
Información sobre
Biodiversidad

- Started in 1998.
- Mexico.

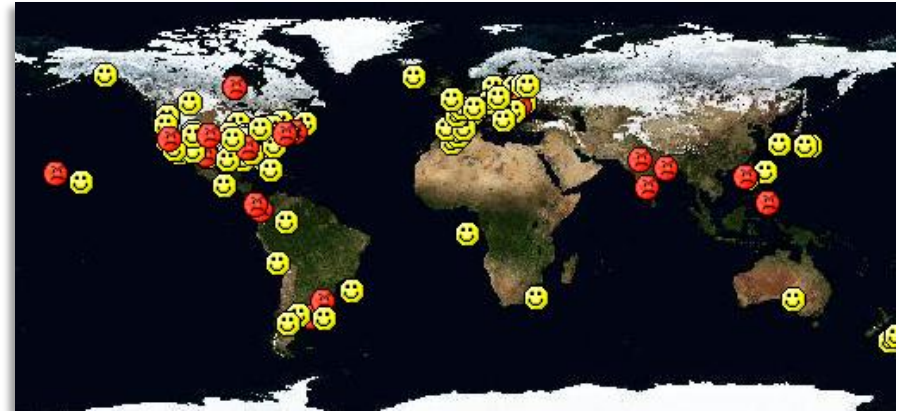


- Started in 2000.
- 9 major herbaria.
- 6 million records (80% databased).

Strategies for data integration

3) Real time distributed queries

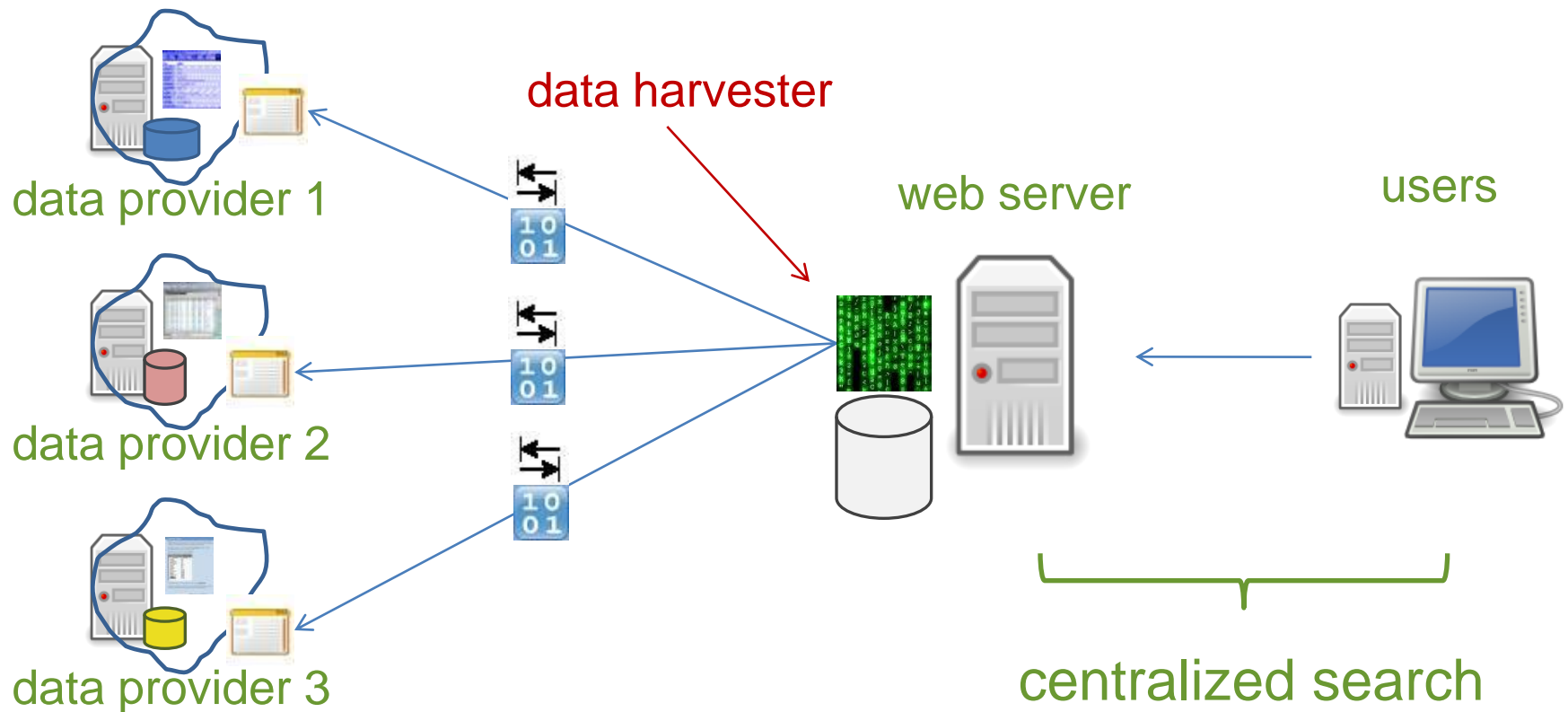
- ✓ Access to current data.
- ✓ Providers have more confidence and sense of control.
- Performance and scalability bottlenecks.
 - Performance limited by the slowest data provider.
 - Servers sometimes down, network problems. When data providers go offline their data become unavailable.



BigDig service monitor

Strategies for data integration

4) Data harvesting



Strategies for data integration

4) Data harvesting

Examples:



- GBIF
- Launched in 2004
- > 200M records



*species*link



Strategies for data integration

4) Data harvesting

- ✓ Good performance.
- It may be necessary to define a common (minimum) field set for storing data in the central database.
- Queries are performed on potentially non current data.
- Difficult to implement if there are many protocols and data standards involved.

Networks architecture evolution

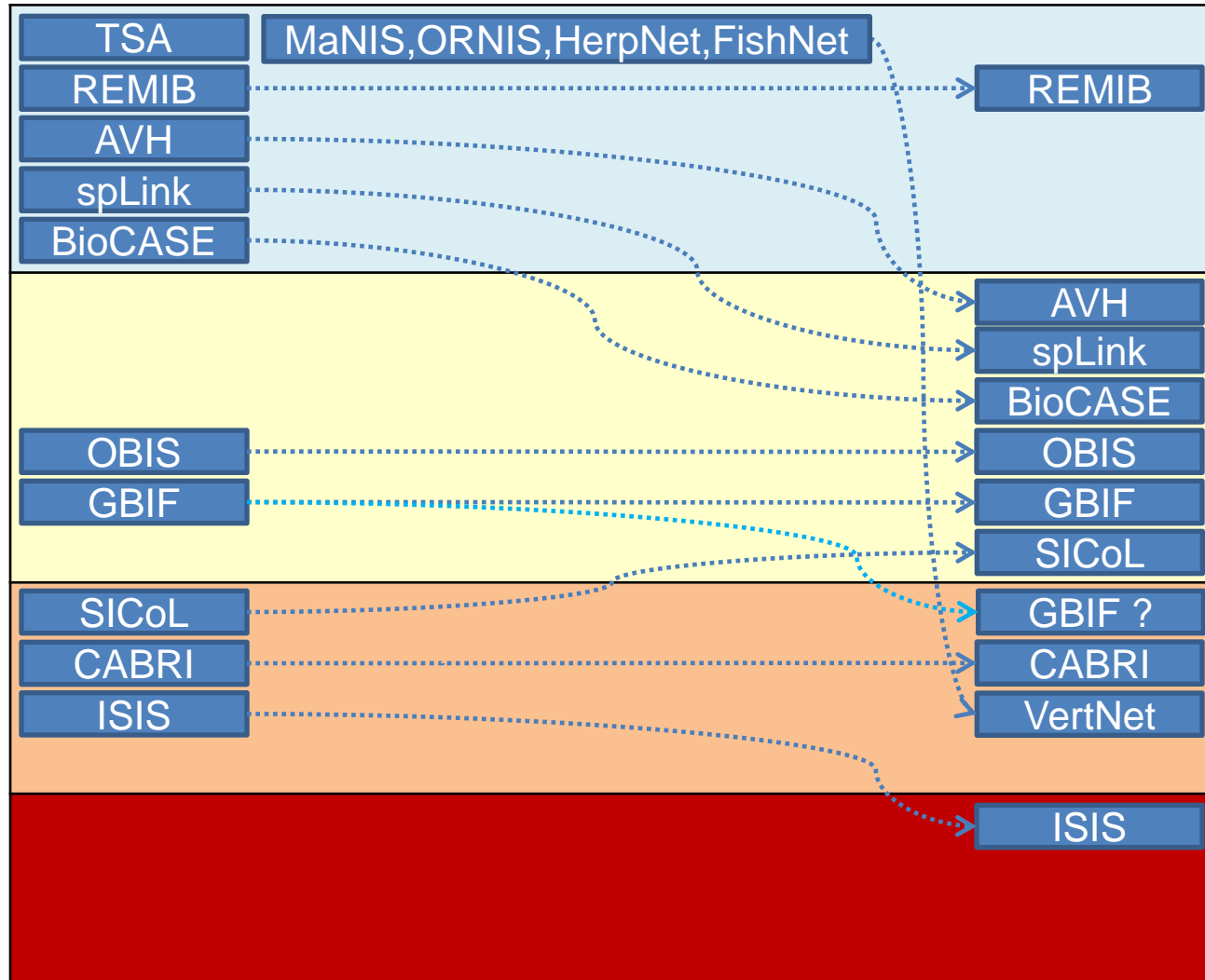
Architecture

Glad to be
of service
(distributed queries)

Let me
harvest you
(data harvesting)

Send me
your data
(data export)

You will be
assimilated!
(single database)



timeline →

About standards



```
<record>
  <DateLastModified>2010-07-01T10:42:01</DateLastModified>
  <InstitutionCode>Fiocruz</InstitutionCode>
  <CollectionCode>CCFF</CollectionCode>
  <CatalogNumber>IOC 0124</CatalogNumber>
  <ScientificName>Saccharomyces cerevisae Hansen</ScientificName>
  <BasisOfRecord>L</BasisOfRecord>
  <Kingdom>Fungo</Kingdom>
  <Genus>Saccharomyces</Genus>
  <Species>cerevisae</Species>
  <ScientificNameAuthor>Hansen</ScientificNameAuthor>
  <YearIdentified>1996</YearIdentified>
  <MonthIdentified>03</MonthIdentified>
  <DayIdentified>07</DayIdentified>
  <Collector>Gilberto Villela</Collector>
  <YearCollected>1943</YearCollected>
  <MonthCollected>01</MonthCollected>
  <DayCollected>01</DayCollected>
</record>
```

Ne

<Scie

ame>

Criteria for choosing a standard

- ✓ Technical features.
- ✓ Flexibility.
- ✓ Stability.
- ✓ Wide adoption.
- ✓ Availability of software & tools.

But there can be political issues above the technical ones!

Standards bodies for Information Technology

ISO, W3C, OGC (Open Geospatial Consortium), OASIS (Organization for the Advancement of Structured Information Standards), among others.

B i o d i v e r s i t y
I n f o r m a t i o n
S t a n d a r d s
T D W G

www.tdwg.org

- International not-for-profit organization.
- Activities started in 1985.
- Mission: Develop, adopt and promote standards to exchange biodiversity data.
- Outcomes are freely available to the public.
- Anyone can participate.

Types of outcomes from TDWG

- Technical Specification:
 - Protocol, service, procedure, format.
- Applicability Statement:
 - How an existing technology can be applied.
- Best Current Practice:
 - Recommended way to proceed in a specific situation.
- Data Standard:
 - Content specification or controlled vocabulary.

Data standards for biological collections

- ABCD
- Darwin Core



ABCD

- **A**ccess to **B**iological **C**ollections **D**ata.
- Defines an XML structure to represent data about objects stored in biological collections.
- Result of TDWG / CODATA (The Committee on Data for Science and Technology) Task Groups.
- Work started in 2000. Official TDWG standard in 2005.
- First used by the European BioCASE network.
- One of the formats supported for sharing data with GBIF.

ABCD

- Comprehensive data representation, therefore complex: 970 terms!!
- Highly structured XML representation.
- Includes specific sections for different types of data:
 - Herbaria.
 - Botanical Gardens.
 - Zoological collections.
 - Plant genetic resources.
 - Palaeontological collections.
 - **Culture collections.**

Darwin Core

- Based on specifications developed by the Dublin Core Metadata Initiative. Can be seen as an extension of it for biodiversity data.
- Its latest version consists of a glossary of terms including definitions, examples, and commentaries, including how terms:
 - are managed
 - can be used
 - can be extended for new purposes
- Designed to minimize the barriers to adoption and to maximize reusability in a variety of contexts.

Darwin Core Terms

Record-level Terms

[dcterms:type](#) | [dcterms:modified](#) | [dcterms:language](#) | [dcterms:rights](#) | [dcterms:rightsHolder](#) | [dcterms:accessRights](#) | [dcterms:bibliographicCitation](#)

[institutionID](#) | [collectionID](#) | [datasetID](#) | [institutionCode](#) | [collectionCode](#) | [datasetName](#) | [ownerInstitutionCode](#) | [basisOfRecord](#) | [informationWithheld](#) | [dataGeneralizations](#) | [dynamicProperties](#)

Occurrence

[occurrenceID](#) | [catalogNumber](#) | [occurrenceDetails](#) | [occurrenceRemarks](#) | [recordNumber](#) | [recordedBy](#) | [individualID](#) | [individualCount](#) | [sex](#) | [lifeStage](#) | [reproductiveCondition](#) | [behavior](#) | [establishmentMeans](#) | [occurrenceStatus](#) | [preparations](#) | [disposition](#) | [otherCatalogNumbers](#) | [previousIdentifications](#) | [associatedMedia](#) | [associatedReferences](#) | [associatedOccurrences](#) | [associatedSequences](#) | [associatedTaxa](#)

Event

[eventID](#) | [samplingProtocol](#) | [samplingEffort](#) | [eventDate](#) | [eventTime](#) | [startDayOfYear](#) | [endDayOfYear](#) | [year](#) | [month](#) | [day](#) | [verbatimEventDate](#) | [habitat](#) | [fieldNumber](#) | [fieldNotes](#) | [eventRemarks](#)

dcterms:Location

[locationID](#) | [higherGeographyID](#) | [higherGeography](#) | [continent](#) | [waterBody](#) | [islandGroup](#) | [island](#) | [country](#) | [countryCode](#) | [stateProvince](#) | [county](#) | [municipality](#) | [locality](#) | [verbatimLocality](#) | [verbatimElevation](#) | [minimumElevationInMeters](#) | [maximumElevationInMeters](#) | [verbatimDepth](#) | [minimumDepthInMeters](#) | [maximumDepthInMeters](#) | [minimumDistanceAboveSurfaceInMeters](#) | [maximumDistanceAboveSurfaceInMeters](#) | [locationAccordingTo](#) | [locationRemarks](#) | [verbatimCoordinates](#) | [verbatimLatitude](#) | [verbatimLongitude](#) | [verbatimCoordinateSystem](#) | [verbatimSRS](#) | [decimalLatitude](#) | [decimalLongitude](#) | [geodeticDatum](#) | [coordinateUncertaintyInMeters](#) | [coordinatePrecision](#) | [pointRadiusSpatialFit](#) | [footprintWKT](#) | [footprintSRS](#) | [footprintSpatialFit](#) | [georeferencedBy](#) | [georeferenceProtocol](#) | [georeferenceSources](#) | [georeferenceVerificationStatus](#) | [georeferenceRemarks](#)

GeologicalContext

[geologicalContextID](#) | [earliestEonOrLowestEonothem](#) | [latestEonOrHighestEonothem](#) | [earliestEraOrLowestErathem](#) | [latestEraOrHighestErathem](#) | [earliestPeriodOrLowestSystem](#) | [latestPeriodOrHighestSystem](#) | [earliestEpochOrLowestSeries](#) | [latestEpochOrHighestSeries](#) | [earliestAgeOrLowestStage](#) | [latestAgeOrHighestStage](#) | [lowestBiostratigraphicZone](#) | [highestBiostratigraphicZone](#) | [lithostratigraphicTerms](#) | [group](#) | [formation](#) | [member](#) | [bed](#)

Identification

[identificationID](#) | [identifiedBy](#) | [dateIdentified](#) | [identificationReferences](#) | [identificationRemarks](#) | [identificationQualifier](#) | [typeStatus](#)

Taxon

[taxonID](#) | [scientificNameID](#) | [acceptedNameUsageID](#) | [parentNameUsageID](#) | [originalNameUsageID](#) | [nameAccordingToID](#) | [namePublishedInID](#) | [taxonConceptID](#) | [scientificName](#) | [acceptedNameUsage](#) | [parentNameUsage](#) | [originalNameUsage](#) | [nameAccordingTo](#) | [namePublishedIn](#) | [higherClassification](#) | [kingdom](#) | [phylum](#) | [class](#) | [order](#) | [family](#) | [genus](#) | [subgenus](#) | [specificEpithet](#) | [infraspecificEpithet](#) | [taxonRank](#) | [verbatimTaxonRank](#) | [scientificNameAuthorship](#) | [vernacularName](#) | [nomenclaturalCode](#) | [taxonomicStatus](#) | [nomenclaturalStatus](#) | [taxonRemarks](#)

Auxiliary Terms

ResourceRelationship

[resourceRelationshipID](#) | [resourceID](#) | [relatedResourceID](#) | [relationshipOfResource](#) | [relationshipAccordingTo](#) | [relationshipEstablishedDate](#) | [relationshipRemarks](#)

MeasurementOrFact

[measurementID](#) | [measurementType](#) | [measurementValue](#) | [measurementAccuracy](#) | [measurementUnit](#) | [measurementDeterminedDate](#) | [measurementDeterminedBy](#) | [measurementMethod](#) | [measurementRemarks](#)

How can I use Darwin Core?

- The standard also includes guidelines about how to use Darwin Core in different contexts, such as:
 - XML.
 - Fielded text files.
 - Tagging content in HTML (under construction).
 - RDF (under construction).
- Provides mechanisms for creating extensions. Examples:
 - Germplasm data (Plant Genetic Resources Network).
 - Annotations to herbarium sheets (under construction by the Botanical Research Institute of Texas).
 - Note: There's an extension for microbial data in an older Darwin Core version. Can be easily upgraded.

ABCD or Darwin Core?

- There are data sharing tools available for both.
- Both are extensible and supported by GBIF.
- Both can be used to share culture collection data.
- ABCD is recommended for sharing quite detailed data about objects stored in biological collections.
- Darwin Core is a simpler alternative for sharing more common biodiversity data.
- Darwin Core can be used to build new representations and structures for biodiversity data (not restricted to XML).

Query protocols

- **DiGIR**
 - **D**istributed **G**eneric **I**nformation **R**etrieval.
 - North American initiative (MaNIS, HerpNet, ORNIS and FishNet networks).
 - Developed to work with DarwinCore.
- **BioCAsE**
 - **B**iological **C**ollection **A**ccess **S**ervice.
 - European initiative: 31 countries (BioCAsE network).
 - Developed to work with ABCD.

Data exchange protocols

- **TAPIR**
 - TDWG Access Protocol for Information Retrieval.
 - Integrates functionality from DiGIR and BioCAsE.
 - Completely independent of the data being exchanged: Works with DarwinCore and ABCD.
 - Official TDWG standard.
 - Tools and documentation available.
- **Other options:** OAI-PMH, WFS, SRU

TAPIR in a nutshell

<http://example.net/mywebservice>

5-

?

:

- Is the service

?

Final recommendations

- Choose from existing standards whenever possible:
 - This can save you considerable time.
 - Will likely avoid interoperability issues in the future.
- Seek compatibility with other initiatives.
 - You can benefit from existing tools.
 - You may get extra functionality/data.
- Data providers are the pillars of every network:
 - Help them improve their data.
 - Ensure that data remain curated at the source.
 - Show them that data sharing promotes citation and usage, giving them credits and visibility.

Thank you !

renato @ cria . org . br