# Darwin Core & Darwin Core Archive

*Christian Svindseth & Dag Endresen, GBIF Norway (Credits: Laura Russell)*

# DATA STANDARDS

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

**ABCD** Access to Biological Collection Data (2005)

**DwC** Darwin Core (2009)

**AC** Audubon Core Multimedia Resources Metadata Schema (2013)

**NCD** Natural Collection Descriptions (Draft)

**EML** Ecological Metadata Language

GBIF

# DATA STANDARDS



HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)

SITUATION: THERE ARE 14 COMPETING STANDARDS.

14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES. YEAH!

SOON: SITUATION: THERE ARE 15 COMPETING STANDARDS.

Important principle: *Re-use of terms from standardized terminologies wherever possible.*

# Darwin Core – a glossary of terms

**Record-level Terms**

dcterms:type | dcterms:modified | dcterms:language | dcterms:rights | dcterms:rightsHolder | dcterms:accessRights | dcterms:bibliographicCitation | dcterms:references | **institutionID** | **collectionID** | **datasetID** | **institutionCode** | **collectionCode** | **datasetName** | ownerInstitutionCode | **basisOfRecord** | informationWithheld | dataGeneralizations | dynamicProperties

**Occurrence**

**occurrenceID** | **catalogNumber** | **recordNumber** | **recordedBy** | individualCount | organismQuantity | organismQuantityType | sex | lifeStage | reproductiveCondition | behavior | establishmentMeans | occurrenceStatus | preparations | disposition | associatedMedia | associatedReferences | associatedSequences | associatedTaxa | otherCatalogNumbers | occurrenceRemarks

**Organism**

**organismID** | organismName | organismScope | assocoatedOccurrences | associatedOrganisms | previousIdentifications | organismRemarks

**MaterialSample** | **LivingSpecimen** | **PreservedSpecimen** | **FossilSpecimen**
**materialSampleID**

**Event** | **HumanObservation** | **MachineObservation**
**eventID** | **parentEventID** | **fieldNumber** | **eventDate** | eventTime | startDayOfYear | endDayOfYear | year | month | day | verbatimEventDate | habitat | samplingProtocol | sampleSizeValue | sampleSizeUnit | samplingEffort | fieldNotes | eventRemarks

**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 | georeferencedDate | 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 | typeStatus | identificationQualifier | dateIdentified | identificationReferences | identificationVerificationStatus | identificationRemarks

**Taxon**

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

**ResourceRelationship** (*Auxiliary Terms*)

**resourceRelationshipID** | **resourceID** | **relatedResourceID** | relationshipOfResource | relationshipAccordingTo | relationshipEstablishedDate | relationshipRemarks

**MeasurementOrFact** (*Auxiliary Terms*)

**measurementID** | measurementType | measurementValue | measurementAccuracy | measurementUnit | measurementDeterminedDate | measurementDeterminedBy | measurementMethod | measurementRemarks

http://rs.tdwg.org/dwc/terms/

**Darwin Core (DwC)** provides a flat list of concepts and terms, expressed using RDF.

- DwC "*extensions*" for the declaration of complementary, additional concepts.
- Reuse concepts from other vocabularies whenever possible, do not re-invent terms.

**Darwin Core Archive (DwC-A)** is a data exchange format based on Darwin Core. DwC-A has a star schema model.

- DwC-A cores, extensions and controlled value vocabularies.
- DwC-A resources always REUSE terms from Darwin Core or other concept vocabularies.

# **Reuse** standardized terms

**Concept vocabularies (skos:conceptSchema, RDF)**

- Darwin Core, Darwin Core "*extensions*", NCD, GNA, Audubon Core (and other concept vocabularies).
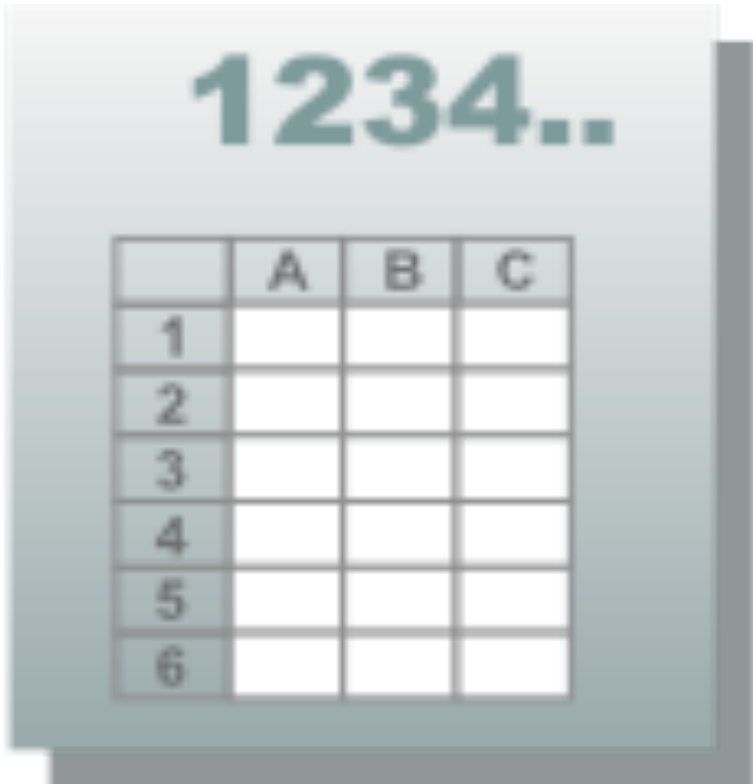
as a basis and foundation for

**Software application schema (XML, XML schema)**

- Darwin Core Archive (DwC-A) controlled terminology and controlled value vocabularies.

- Resources such as the DwC-A controlled term and value vocabularies REUSE terms (by URI) from a concept vocabulary.

# SIMPLE DARWIN CORE

**Simple-DwC** is a specification for one particular way to use the Darwin Core terms - to share data about taxa and their occurrences in a flat and simply structured way - and is probably what is meant if someone suggests to "*format your data according to the Darwin Core*".

GBIF

# DATA NORMALIZATION

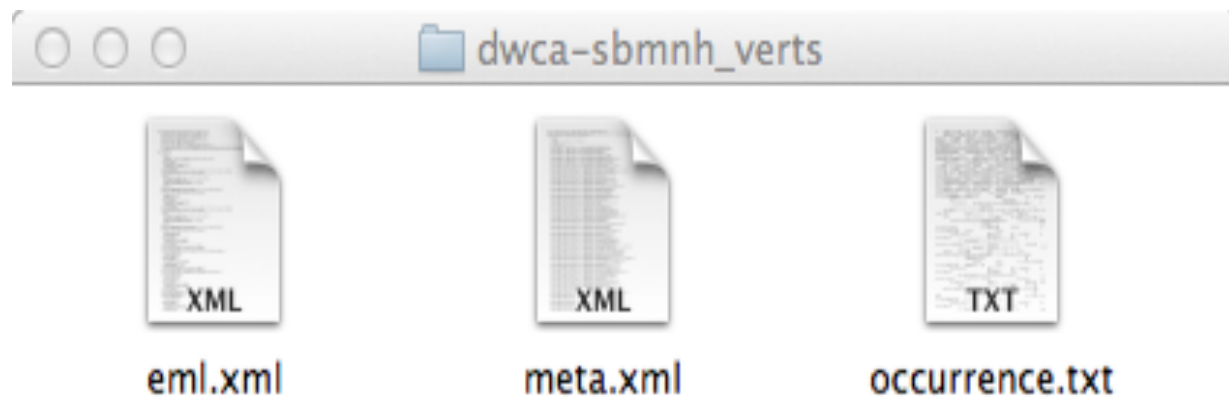What is data normalization?

Reasons to normalize a database

Normal forms

"*Data normalization* is the process of reducing data to its canonical form. For instance, Database normalization is the process of organizing the fields and tables of a relational database to **minimize redundancy and dependency**" (Wikipedia).

"*Denormalization* is the process of attempting to **optimize the read performance** of a database by adding redundant data or by grouping data" (Wikipedia).
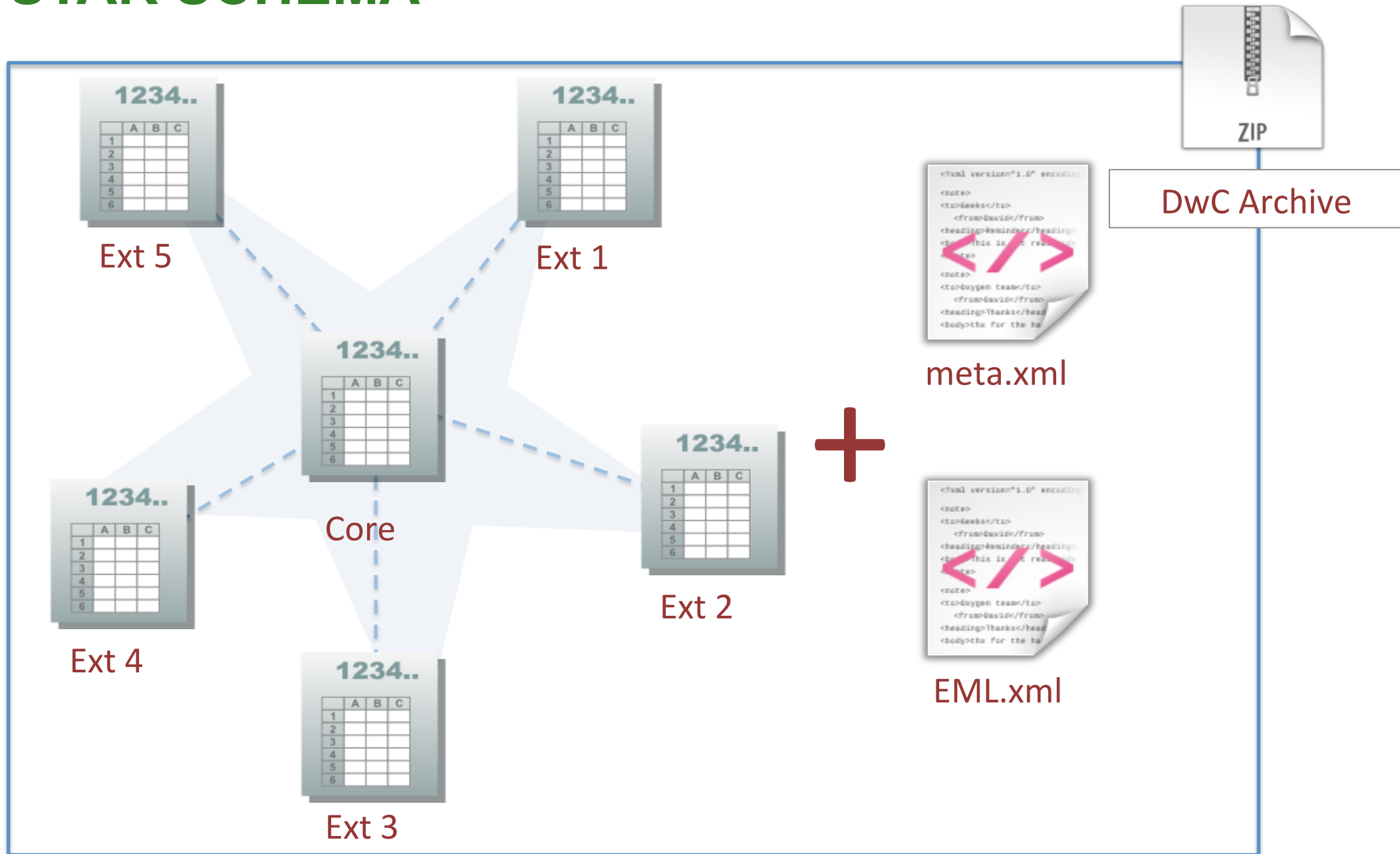
GBIF

# DARWIN CORE ARCHIVE

A Darwin Core Archive (DwC-A) is the text representation of data formatted to Darwin Core.

A DwC-A is a compressed file containing a <u>minimum</u> of three files.

# STAR SCHEMA

# MAPPING CORES

## Taxon core

The category of information pertaining to taxonomic names, taxon name usages, or taxon concepts. Updated in April 2015, 43 terms.

This version removes terms dcterms:source and dcterms:rights, and adds dcterms:license.

## Occurrence core

The category of information pertaining to evidence of an occurrence in nature, in a collection, or in a dataset (specimen, observation, etc.). Updated in July 2015, 169 terms.

This version removes terms dcterms:source, dcterms:rights, dwc:individualID, dwc:occurrenceDetails, and adds dcterms:license, dwc:organismQuantity, dwc:organismQuantityType, dwc:organismID, dwc:organismName, dwc:organismScope, dwc:associatedOrganisms, dwc:organismRemarks, dwc:parentEventID, dwc:sampleSizeValue, dwc:sampleSizeUnit.

## Event core

The category of information pertaining to a sampling event. Issued 29 May 2015, 95 terms.
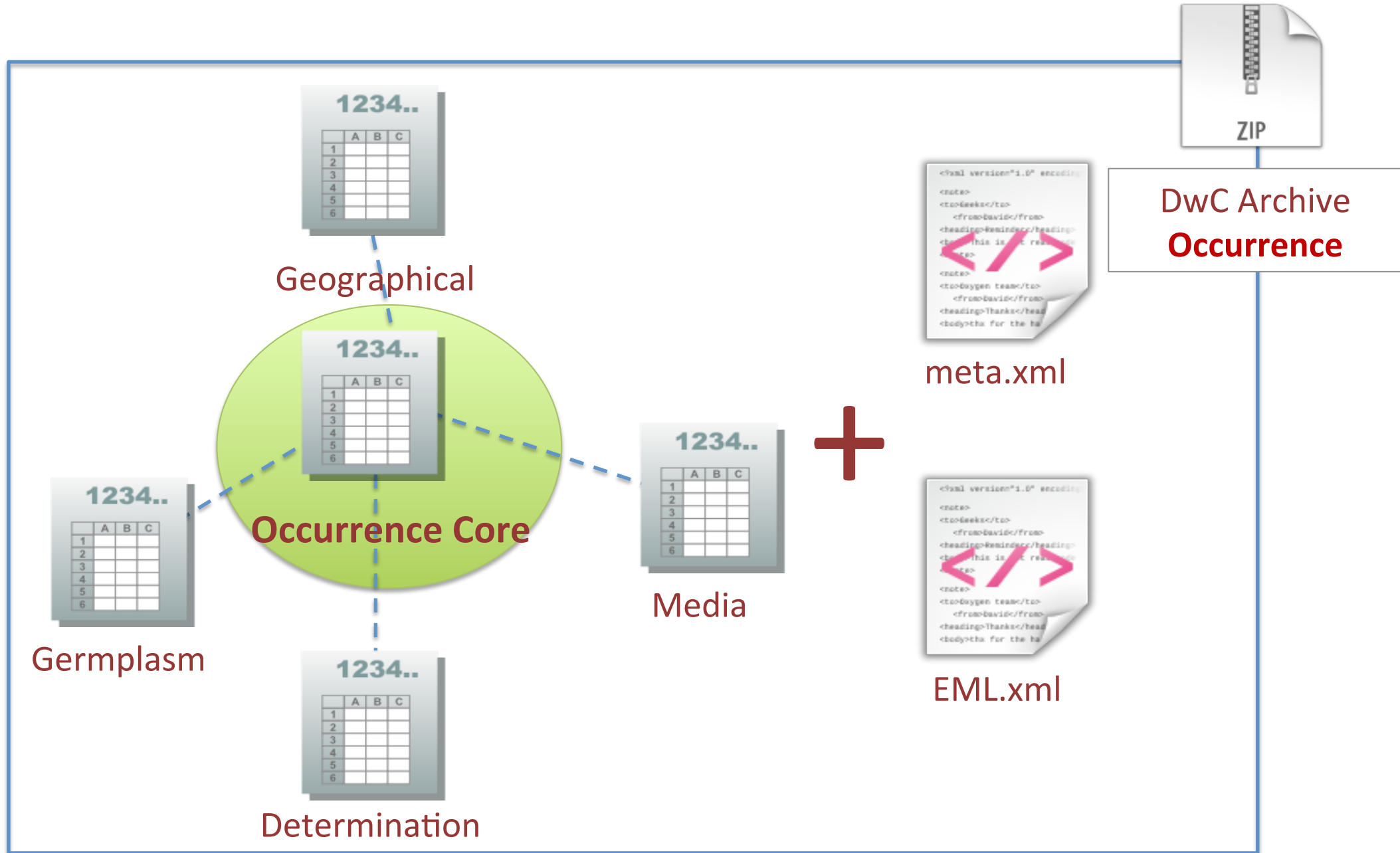
GBIF

# EXTENSIONS

Darwin Core does not provide terms for every possible type of data.

- 22 registered extensions (to the Darwin Core Archive)
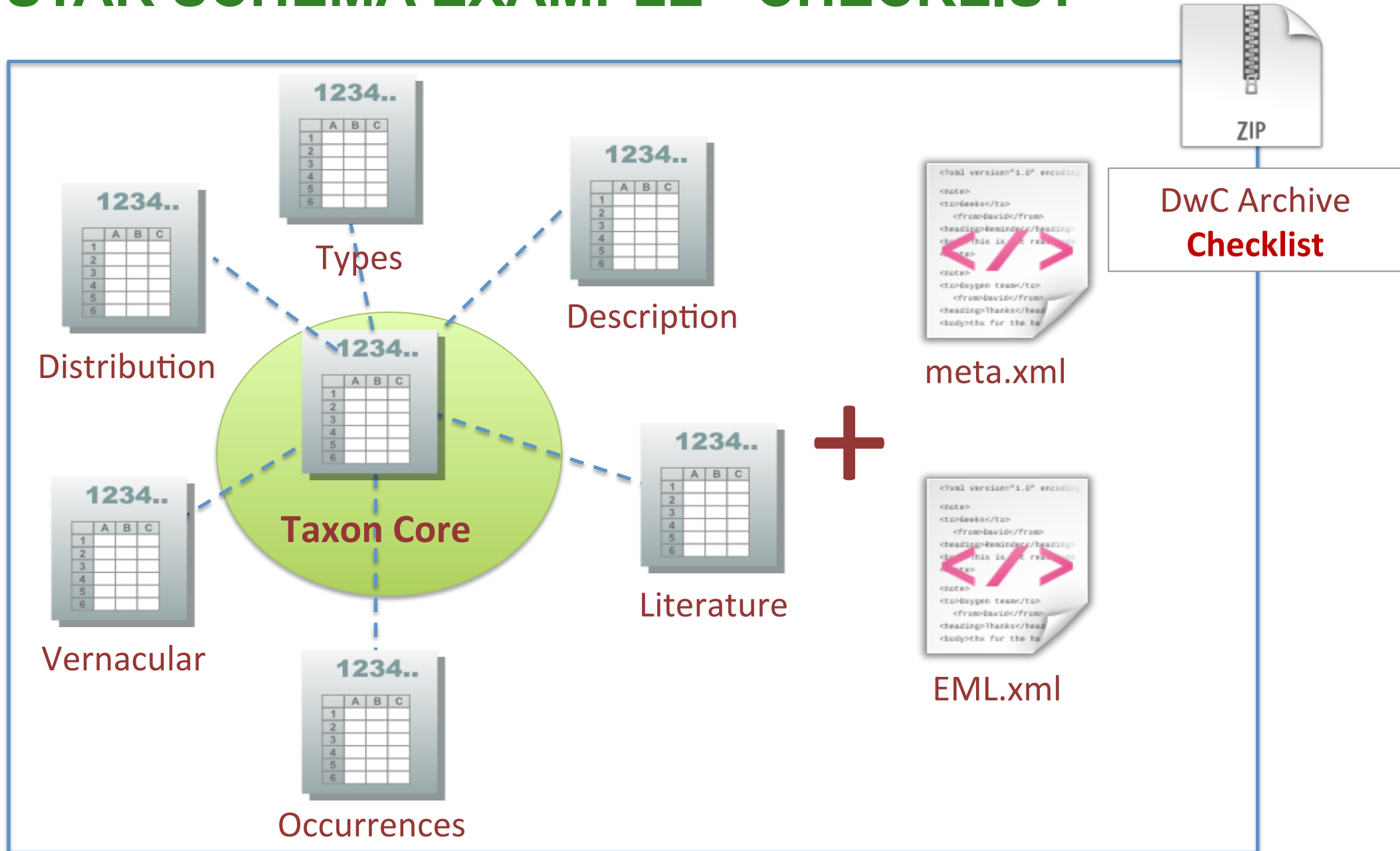- 25 under development

Examples

- Audubon Media Description (aka Audubon Core)
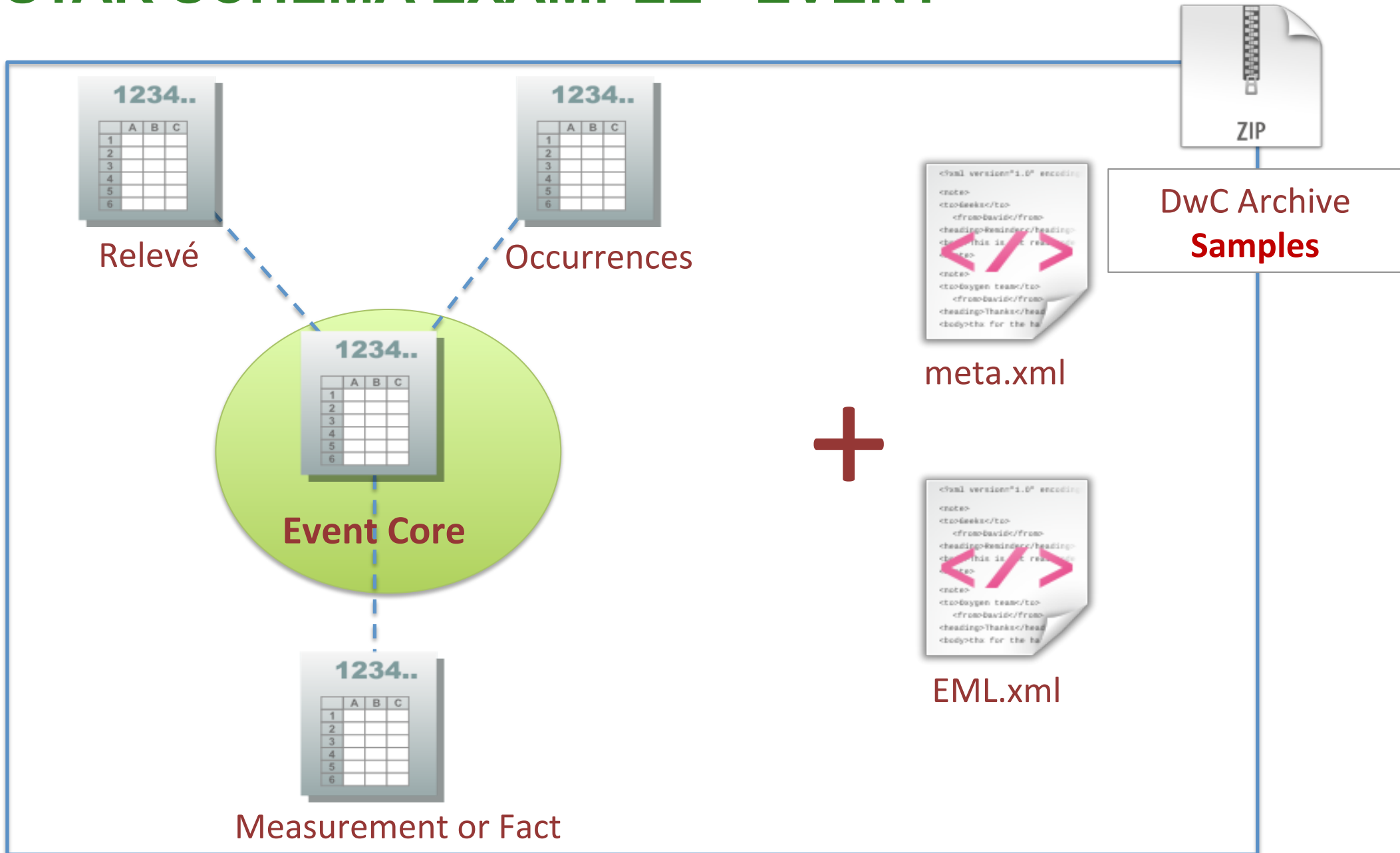- Darwin Core Identification History
- Darwin Core Measurement or Facts

GBIF

# STAR SCHEMA EXAMPLE - OCCURRENCE



Geographical

Occurrence Core

Germplasm

Media

Determination

meta.xml

EML.xml

DwC Archive
**Occurrence**

GBIF

# STAR SCHEMA EXAMPLE - CHECKLIST



Types

Description

Distribution

Taxon Core

Literature

meta.xml

Vernacular

Occurrences

EML.xml

ZIP

DwC Archive
**Checklist**

**+**

GBIF

# STAR SCHEMA EXAMPLE - EVENT



Relevé

Occurrences

Event Core

Measurement or Fact

meta.xml

EML.xml

+

ZIP

DwC Archive
**Samples**

GBIF

# Darwin Core & Darwin Core Archive

*Christian Svindseth & Dag Endresen, GBIF Norway (Credits: Laura Russell)*