Stable Identifiers for Natural History Collections

Anton Güntsch

Freie Universität Berlin

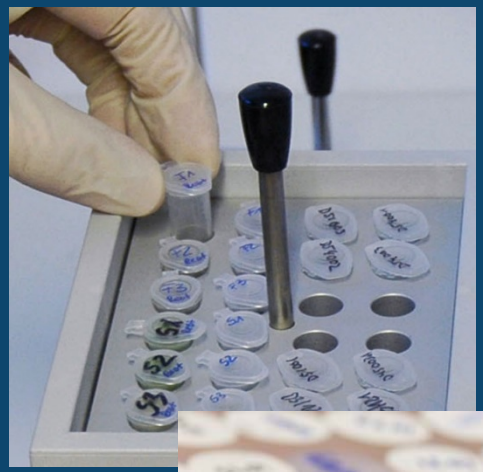
BG BM Botanischer Garten & Botanisches Museum Berlin

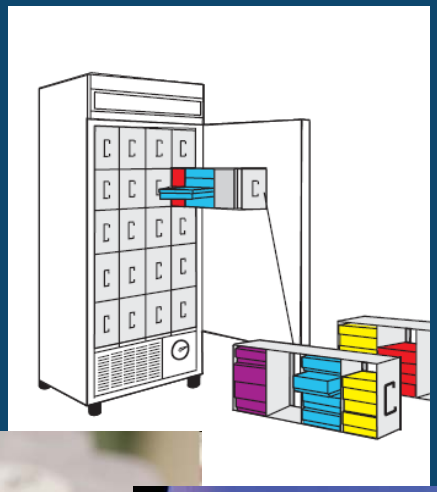
Natural History Collections

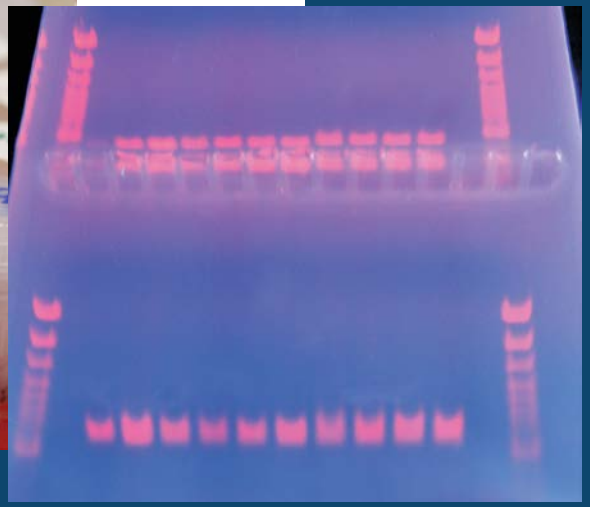
Natural History Collections

Natural History Collections

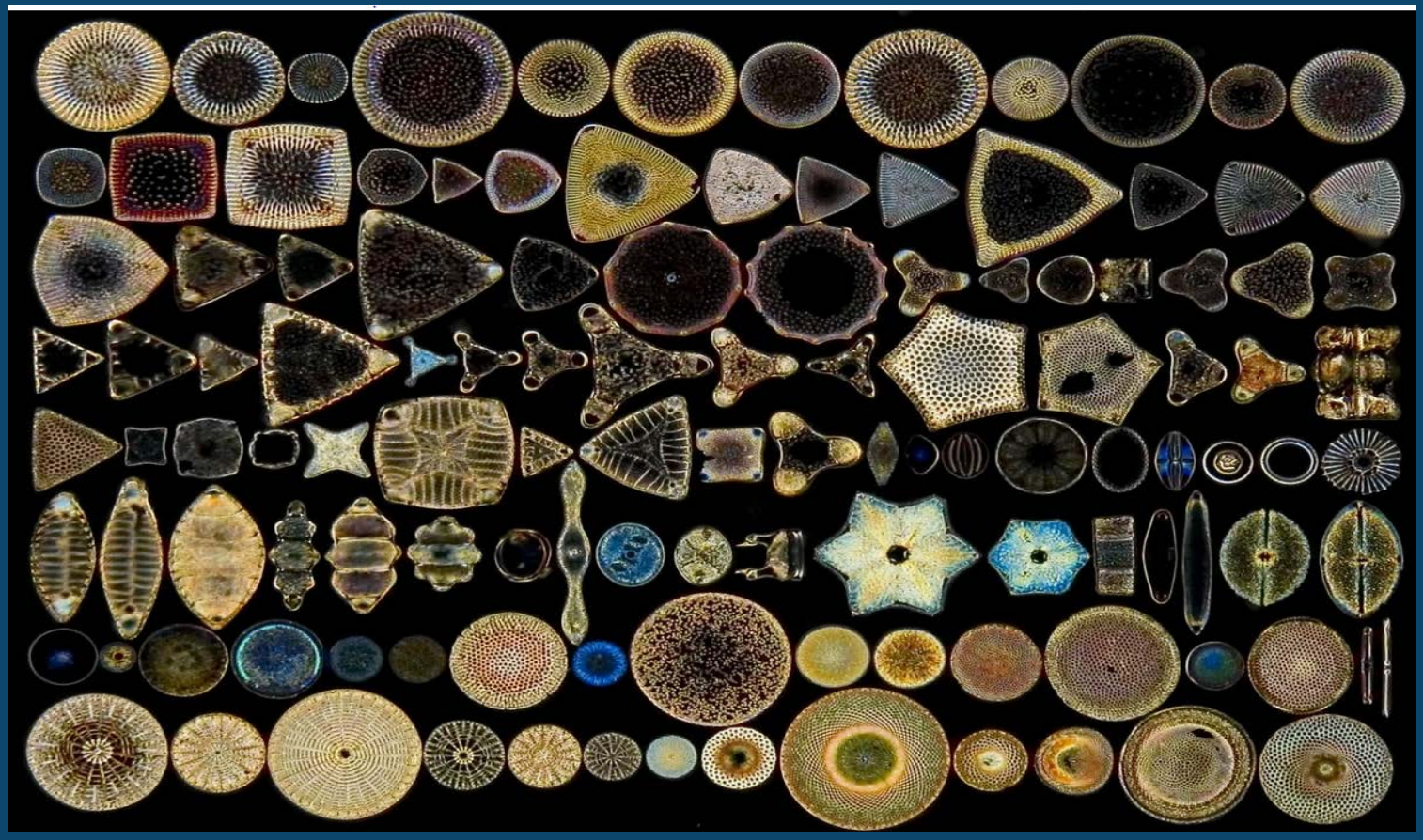
Natural History Collections







Natural History Collections



© Matthias Burba, 2009

A physical database of the natural world

Provides the basis for

- morphological studies
- species inventories
- studies on biodiversity changes
- distribution modelling
- effect of climate change in biodiversity
- prediction of invasion of species
- …

How many?

How many?

2-3 Billion **physical** objects*

*Duckworth WD, Genoways HH, Rose CL. Preserving natural science collections: Chronicle of our environmental heritage. National Institute for the Conservation of Cultural Property, Washington, D.C.; 1993. 140 pp

How many digitised?

How many digitised?

- We don‘t really know.

How many digitised?

- We don‘t really know.
- The Global Biodiversity Information Facility (GBIF) provides access to 120 Million records.

→ Perhaps between 5% and 10% digitised.

Digitisation rate is rapidly growing

Digitisation rate is rapidly growing

- Paris Herbarium: 6 million specimens went online between 2008 and 2012
- Similar activities in the Netherlands, Finland, Norway, US, …

Identifiers (traditionally)

Used as a mechanism for referencing specimens in publications, catalogues, etc.

Iresine angustifolia Euphrasén, Beskr. Svenska Vestindiska, 165. 1795.

aroma de medianoche, chivo, coyuntura negra, siete pellejos, tacuquelite

Distribución. – México, Guatemala, El Salvador y Honduras hasta Panamá y hasta Brasil y Perú; Las Antillas.

Muestra(s). – AHUACHAPÁN: F. Chinchilla & E. Sandoval s.n. [ISB00153] (B, F, LAGU, MO); J. M. Rosales 399, 494 (B, BM, F, LAGU, MO). LA LIBERTAD: R. Aparicio & R. Rivera 138 (B, F, LAGU); R. Cruz 8 (LAGU), s.n. [WB-00504] (B, LAGU, MO). CHALATENANGO: R. Villacorta & L. Lara 2532 (B, F, LAGU, MO).

Rep. para El Salvador. – Linares 2005: 114. Berendsohn 1991: 45. Standley & Steyermark 1946, Fl. Guat. [illegible]167. Standley & Calderón 1925: 75.

Otras Ref's. – Borsch 2001, Fl. Nic. 1: 77. Burger 1983, Fl. Costaric. #64: 173. Duke 1961, Fl. Pan. #5[illegible]33.

Ilustración. – Burger 1983, Fl. Costaric. #64: 171, fig. 30.

PLANTAS DE EL SALVADOR

Amaranthaceae

***Iresine angustifolia* Euphrasen**

Depto Chalatenango. San Ignacio,El Rosario. 1100 m. 14°23'N, 89°2.
Creciendo sobre rocas secas a orilla del río, suelos arcillosos. Arbol pequeño de unos 5m de altura, corteza clara. Hojas lanceoladas de c verde oscuro. Flores de color blanco. [RV/rv] No. Herb. LAGU: R 02532

Det. R.Villacorta 4.97

Herbaria: F MO B LAGU

Leg.: R.Villacorta 2532 con L.Lara 4.4.1997

Identifiers (traditionally)

- Informal syntax

Identifiers (traditional)

- Informal syntax
- Uniqueness not enforced

Identifiers (traditional)

- Informal syntax
- Uniqueness not enforced
- No mechanism for linking up Metadata and images

We need a new system

- Clear syntax and resolving mechanisms

We need a new system

- Clear syntax and resolving mechanisms
- Enforcing global uniqueness

We need a new system

- Clear syntax and resolving mechanisms
- Enforcing global uniqueness
- Automated access to Metadata

We need a new system

- Clear syntax and resolving mechanisms
- Enforcing global uniqueness
- Autmated linking to Metadata
- And most important: **accepted**

A long debate in the community starting in the year 2000

UUID

DOI

LSIDs

PURL

Handle System

HTTP URIs

And the winner was …

And the winner is

Biodiversity Information Standards TDWG

Biodiversity Information Standards (TDWG)

http://www.tdwg.org

TDWG Life Sciences Identifiers (LSID) Applicability Statement

Date:
17-Dec-2007

Status:
TDWG Draft Standard

Why LSIDs were not successfull

```
urn:lsid:authority:ns:obj:rev
```

Why LSIDs were not successfull

`urn:lsid:authority:ns:obj:rev`

- Complex syntax and rules

Why LSIDs were not successfull

`urn:lsid:authority:ns:obj:rev`

- Complex syntax and rules
- Access to Metadata requires additional services

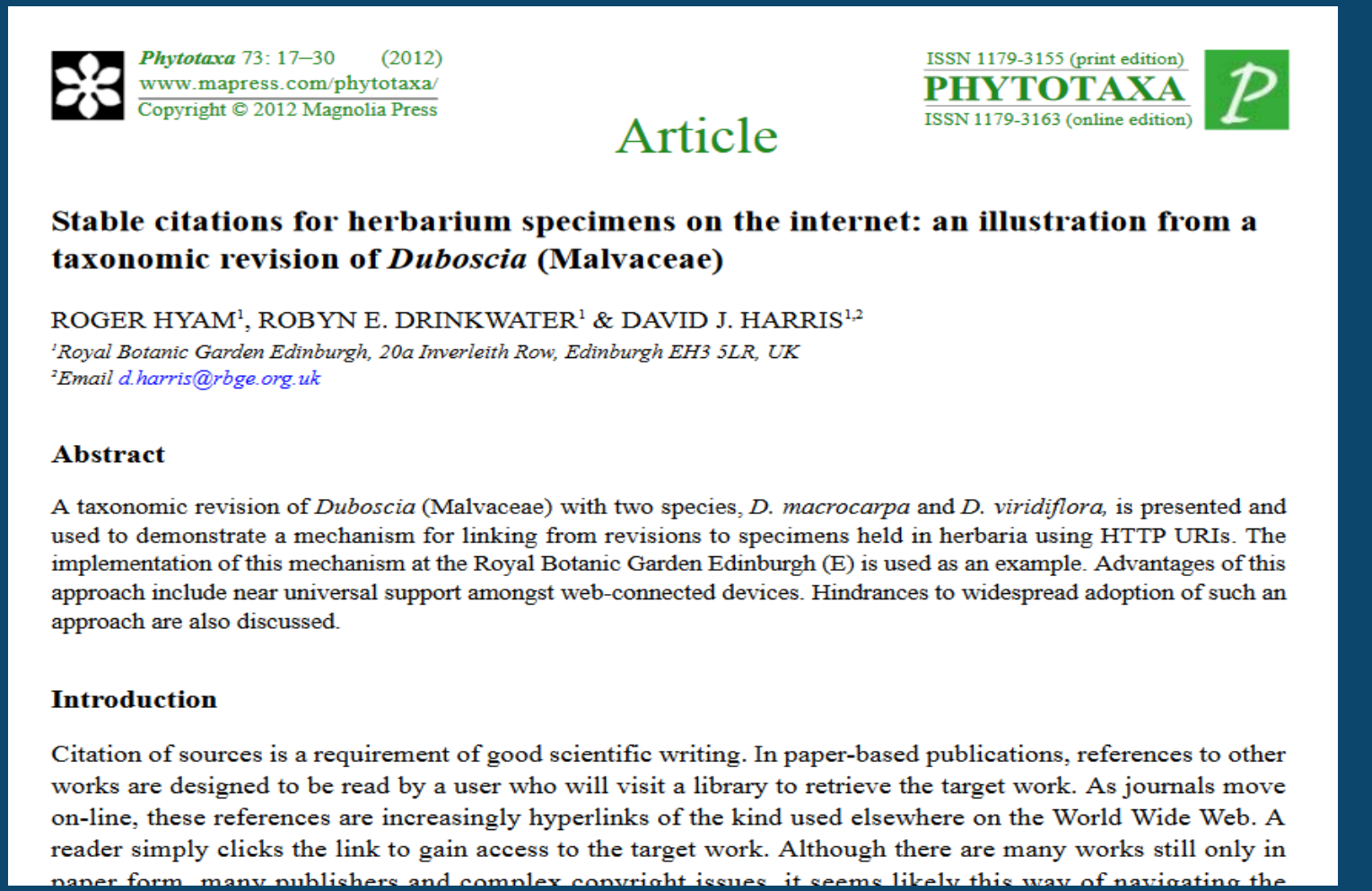
Why LSIDs were not successfull

`urn:lsid:authority:ns:obj:rev`

- Complex syntax and rules
- Access to Metadata requires additional services

→ no conistent implementation in the community

HTTP URI-based specimen identifiers



Phytotaxa 73: 17–30 (2012)
www.mapress.com/phytotaxa/

Article

ISSN 1179-3155 (print edition)
PHYTOTAXA
ISSN 1179-3163 (online edition)

Stable citations for herbarium specimens on the internet: an illustration from a taxonomic revision of *Duboscia* (Malvaceae)

ROGER HYAM[1], ROBYN E. DRINKWATER[1] & DAVID J. HARRIS[1,2]
[1]*Royal Botanic Garden Edinburgh, 20a Inverleith Row, Edinburgh EH3 5LR, UK*
[2]*Email d.harris@rbge.org.uk*

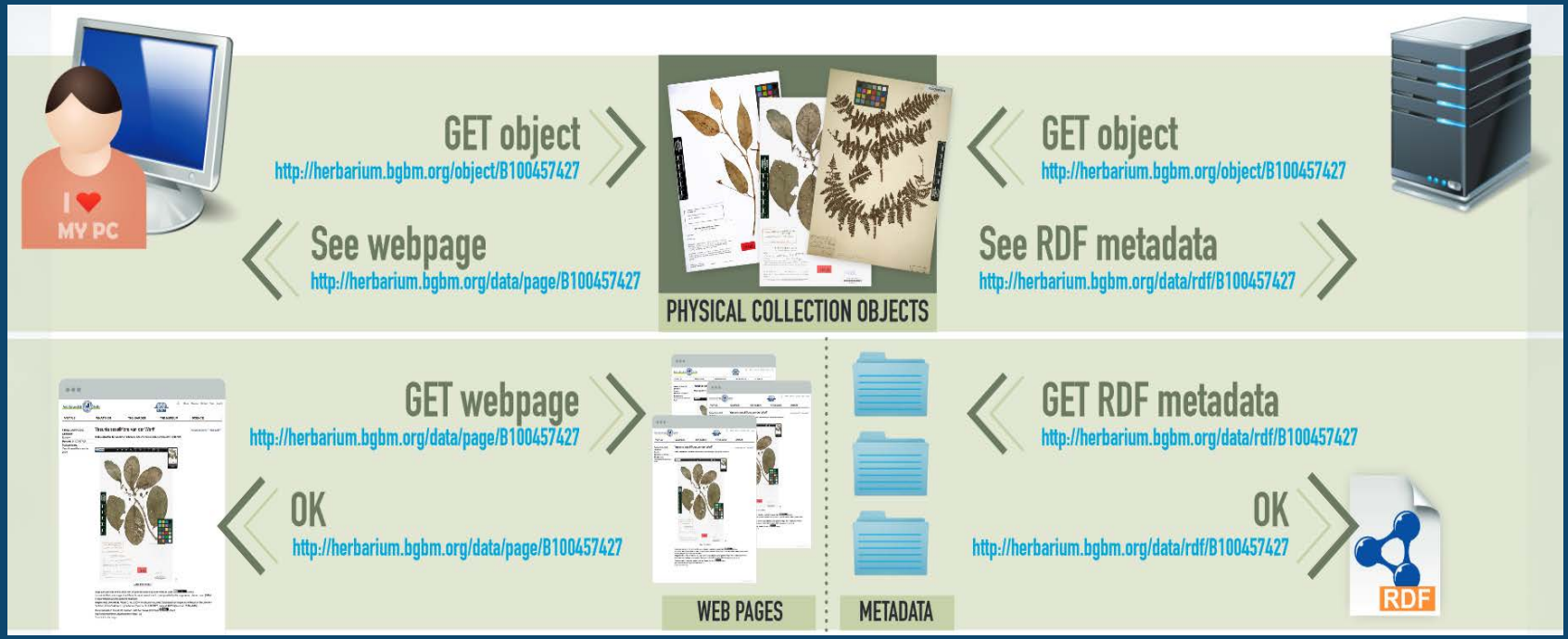
Abstract

A taxonomic revision of *Duboscia* (Malvaceae) with two species, *D. macrocarpa* and *D. viridiflora,* is presented and used to demonstrate a mechanism for linking from revisions to specimens held in herbaria using HTTP URIs. The implementation of this mechanism at the Royal Botanic Garden Edinburgh (E) is used as an example. Advantages of this approach include near universal support amongst web-connected devices. Hindrances to widespread adoption of such an approach are also discussed.

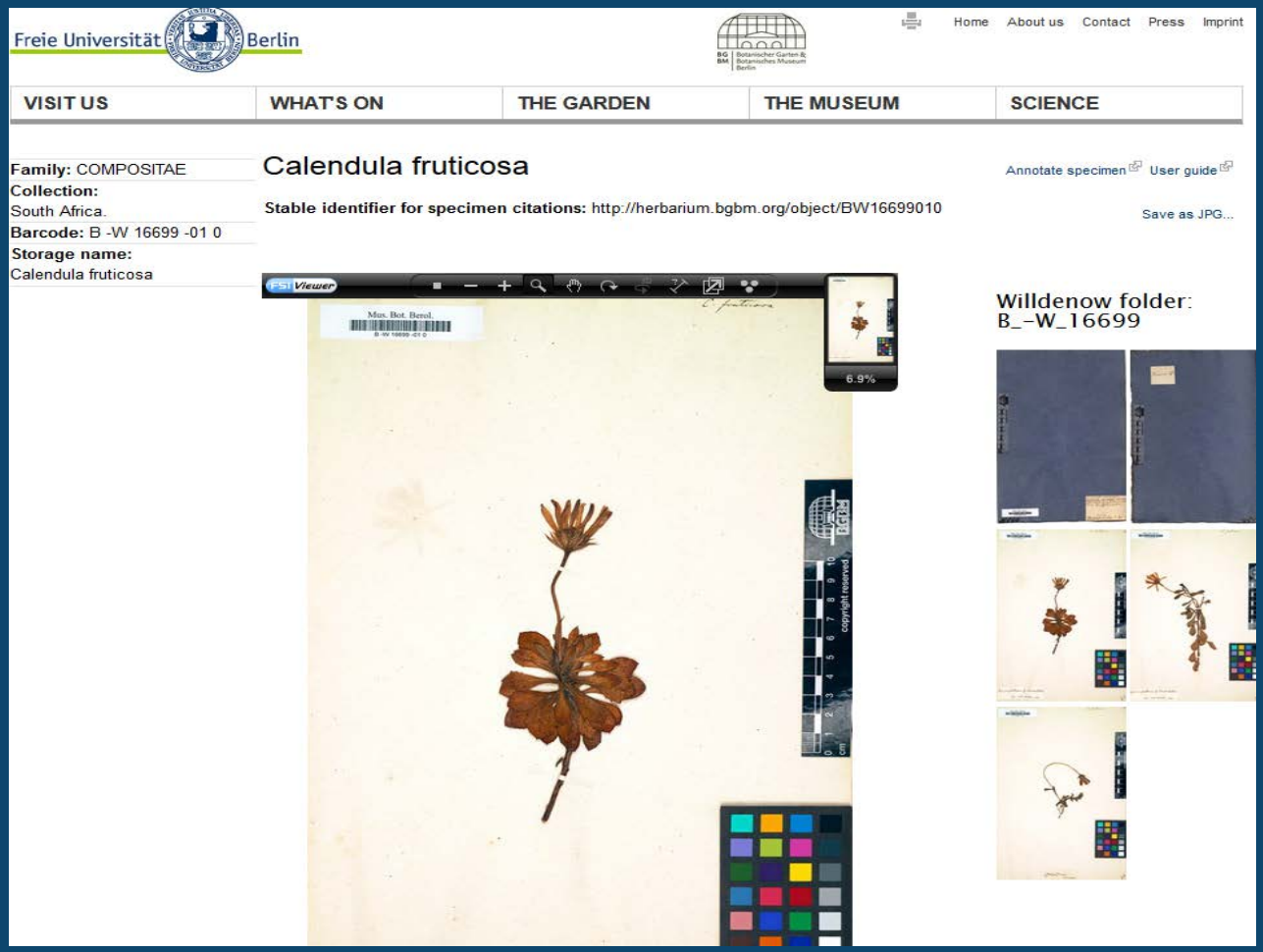
Introduction

Citation of sources is a requirement of good scientific writing. In paper-based publications, references to other works are designed to be read by a user who will visit a library to retrieve the target work. As journals move on-line, these references are increasingly hyperlinks of the kind used elsewhere on the World Wide Web. A reader simply clicks the link to gain access to the target work. Although there are many works still only in paper form, many publishers and complex copyright issues, it seems likely this way of navigating the

HTTP URI-based specimen identifiers



Redirect to human-readable Metadata



Redirect to machine-readable Metadata

http://herbarium.bgbm.org/object/BW16699010 (EasyRdf_Resource)

→ dc:title → "Calendula fruticosa"

→ dc:description → "A herbarium specimen of Calendula fruticosa"

→ dc:type → "Specimen"

→ dc:publisher → "BGBM"

→ http://rs.tdwg.org/dwc/terms/SampleID → "http://herbarium.bgbm.org/object/BW16699010"

→ dc:modified → "2012-12-19 17:10:26.5700000"

→ http://rs.tdwg.org/dwc/terms/basisOfRecord → "Specimen"

→ http://rs.tdwg.org/dwc/terms/CollectionCode → "B"

→ http://rs.tdwg.org/dwc/terms/CatalogNumber → "B -W 16699 -01 0"

→ http://rs.tdwg.org/dwc/terms/scientificName → "Calendula fruticosa"

→ http://rs.tdwg.org/dwc/terms/previousIdentifications → "Calendula fruticosa"

→ http://rs.tdwg.org/dwc/terms/family → "COMPOSITAE"

→ http://rs.tdwg.org/dwc/terms/genus → "Calendula"

→ http://rs.tdwg.org/dwc/terms/specificEpithet → "fruticosa"

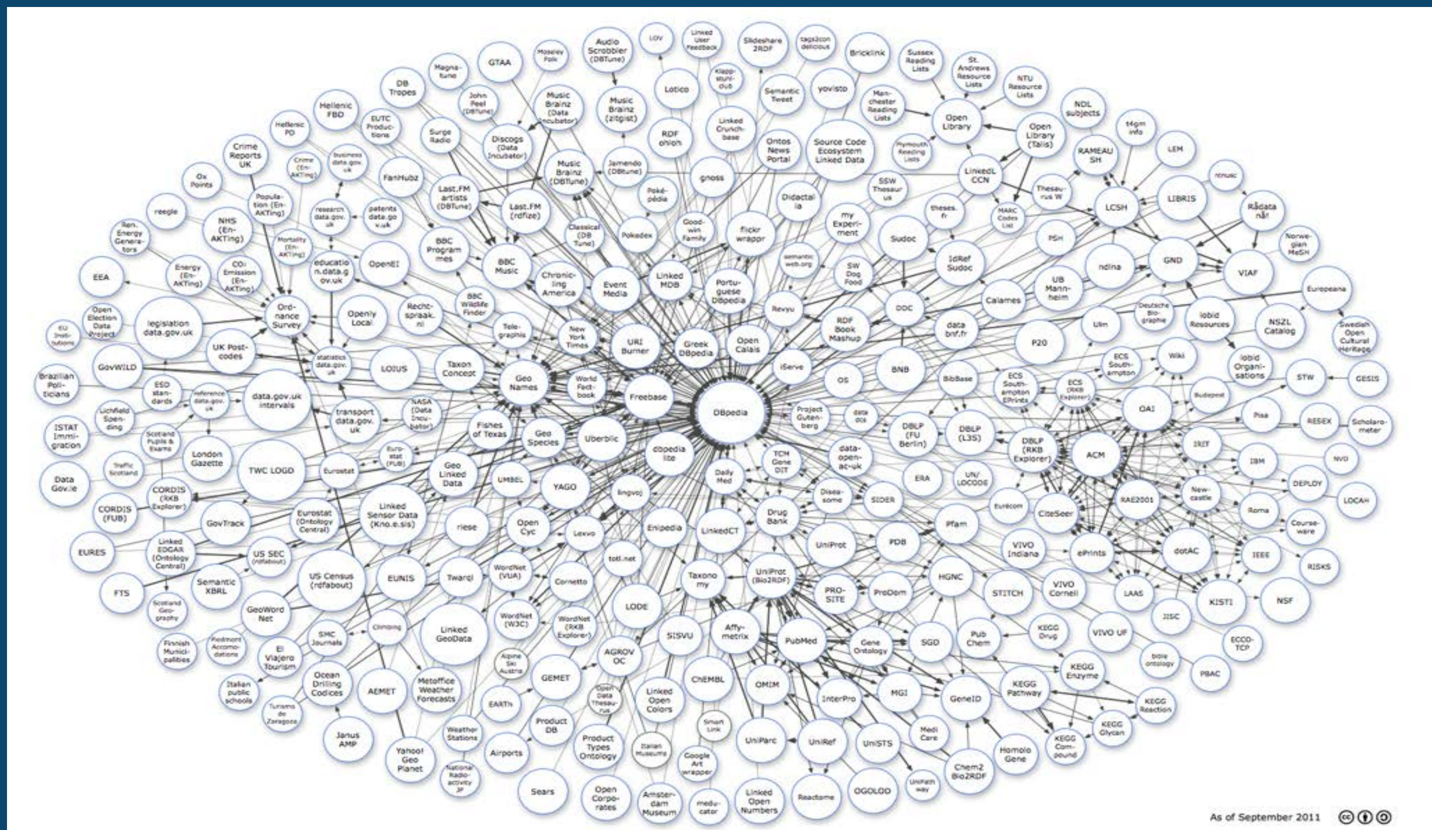
→ http://rs.tdwg.org/dwc/terms/higherGeography → "South Africa"

→ http://rs.tdwg.org/dwc/terms/countryCode → "South Africa"

→ http://rs.tdwg.org/dwc/terms/Locality → "South Africa."

→ http://rs.tdwg.org/dwc/terms/associatedMedia → http://ww2.bgbm.org/herbarium/images/B/-W/16/69/B_-W_16699 -01 0.jpg

Compliance to Linked Open Data (LOD) and the semantic web



Freie Universität Berlin

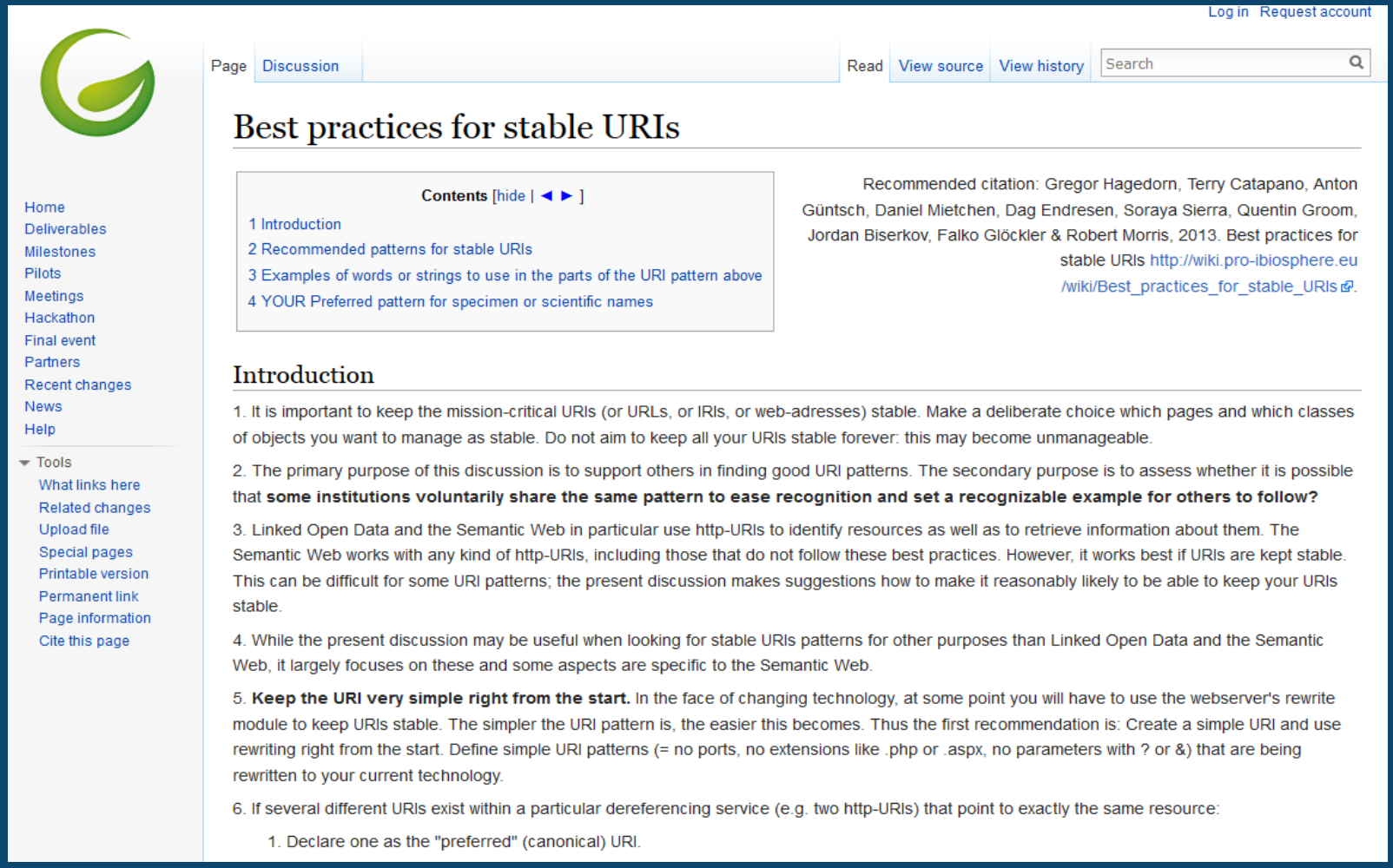
BG BM Botanischer Garten & Botanisches Museum Berlin

2013: HTTP URIs discussed in CETAF

2013: HTTP URIs discussed in CETAF

Surprise ☺ Edinburgh (RBGE), Berlin (BGBM), Berlin (MfN), MNHN (Paris) decide to have stable HTTP-URI-based identifiers implemented within 3 months.

Joint development of best practices



Log in Request account

Page Discussion Read View source View history Search

Best practices for stable URIs

Home
Deliverables
Milestones
Pilots
Meetings
Hackathon
Final event
Partners
Recent changes
News
Help

Tools
What links here
Related changes
Upload file
Special pages
Printable version
Permanent link
Page information
Cite this page

Contents [hide | ◄ ►]

1 Introduction
2 Recommended patterns for stable URIs
3 Examples of words or strings to use in the parts of the URI pattern above
4 YOUR Preferred pattern for specimen or scientific names

Recommended citation: Gregor Hagedorn, Terry Catapano, Anton Güntsch, Daniel Mietchen, Dag Endresen, Soraya Sierra, Quentin Groom, Jordan Biserkov, Falko Glöckler & Robert Morris, 2013. Best practices for stable URIs http://wiki.pro-ibiosphere.eu/wiki/Best_practices_for_stable_URIs.

Introduction

1. It is important to keep the mission-critical URIs (or URLs, or IRIs, or web-adresses) stable. Make a deliberate choice which pages and which classes of objects you want to manage as stable. Do not aim to keep all your URIs stable forever: this may become unmanageable.

2. The primary purpose of this discussion is to support others in finding good URI patterns. The secondary purpose is to assess whether it is possible that **some institutions voluntarily share the same pattern to ease recognition and set a recognizable example for others to follow?**

3. Linked Open Data and the Semantic Web in particular use http-URIs to identify resources as well as to retrieve information about them. The Semantic Web works with any kind of http-URIs, including those that do not follow these best practices. However, it works best if URIs are kept stable. This can be difficult for some URI patterns; the present discussion makes suggestions how to make it reasonably likely to be able to keep your URIs stable.

4. While the present discussion may be useful when looking for stable URIs patterns for other purposes than Linked Open Data and the Semantic Web, it largely focuses on these and some aspects are specific to the Semantic Web.

5. **Keep the URI very simple right from the start.** In the face of changing technology, at some point you will have to use the webserver's rewrite module to keep URIs stable. The simpler the URI pattern is, the easier this becomes. Thus the first recommendation is: Create a simple URI and use rewriting right from the start. Define simple URI patterns (= no ports, no extensions like .php or .aspx, no parameters with ? or &) that are being rewritten to your current technology.

6. If several different URIs exist within a particular dereferencing service (e.g. two http-URIs) that point to exactly the same resource:

    1. Declare one as the "preferred" (canonical) URI.

To date, stable IDs implemented by

Botanischer Garten und Botanisches Museum Berlin-Dahlem; Finnish Museum of Natural History, Helsinki; Museum für Naturkunde, Berlin; Muséum national d'histoire naturelle, Paris; Naturalis Biodiversity Center, Leiden; The Natural History Museum, London; Natural History Museum - University of Oslo; Royal Botanic Garden Edinburgh; Royal Botanic Garden Kew; Staatliches Museum für Naturkunde Stuttgart; Staatliche Naturwissenschaftliche Sammlungen Bayerns; Zoologisches Forschungsmuseum Alexander Koenig, Bonn

… and

Harvard University Herbaria; Harvard Museum of Comparative Zoology

Why does it work?

Why does it work?

1) HTTP-URIs are (technically) easy to implement.

Why does it work?

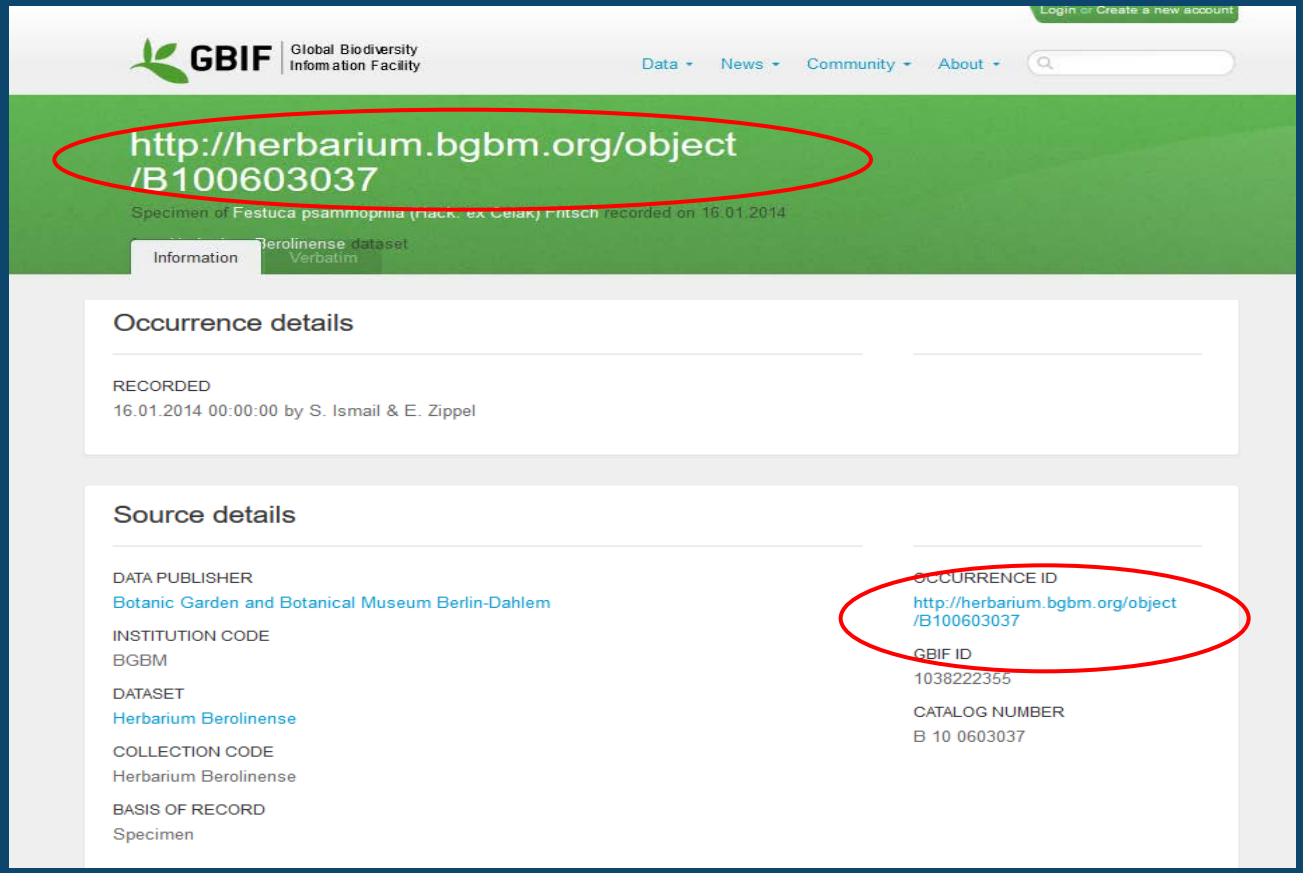
1) HTTP-URIs are (technically) easy to implement.
2) HTTP-URIs are easy to use.

Why does it work?

1) HTTP-URIs are (technically) easy to implement.
2) HTTP-URIs are easy to use.
3) There is no binding syntax. Institutions can decide themselves.

Identifiers in use – institutional portals

Identifiers in use – international data portals



Identifiers in use – publications

A single collection was cited in the protologue of the name, but none of the three duplicates mentioned above carry the species name nor were they annotated by Hector Léveillé. Lauener (1965) indicated that the type is at E, but he did not specify which of the two sheets there is the type, and therefore a second-step lectotypification is provided here.

3. ***Cardamine circaeoides*** Hook.f. & Thomson, J. Proc. Linn. Soc., Bot. 5: 144. 1861 (http://ipni.org/urn:lsid:ipni.org:names:280232-1:1.1.2.1.1.3). Described from: [INDIA] "In Himalaya orientali temperata, Sikkim interiore, sylvis, alt. 5000-7000 ped. ! J. D. H. (v.v.)". Lectotype (designated here, or perhaps holotype): [Label 1]: [INDIA] "Hab. Sikkim, Regio temp., *J. D. H[ooker]*"; [Label 2]: [INDIA] "295 Hab. wet wood, Gohsun, Sikkim, 5000 ft." – K! (K000077050 [http://specimens.kew.org/herbarium/K000077050]); Doubtful isolectotypes – B! (B 10 0386925 [http://herbarium.bgbm.org/object/B100386925]), P! (P00747534 [http://coldb.mnhn.fr/catalognumber/mnhn/p/p00747534]).

The K specimen above was collected at an elevation of 5,000 ft, which is in agreement with the protologue, whereas labels of the B and P sheets indicate the elevation of 6,000–10,000 ft. It is questionable whether the three specimens above were collected from the same area, and that is why we feel that the B and P specimens are doubtful isolectotypes.

PhytoKeys

Identifiers in use – automated testing

CETAF URI TESTER | IMPLEMENTERS | DOCUMENTATION | CONTACT

CETAF Specimen URI Tester

Enter a URI in the box below and click 'Check Now' or click for example.

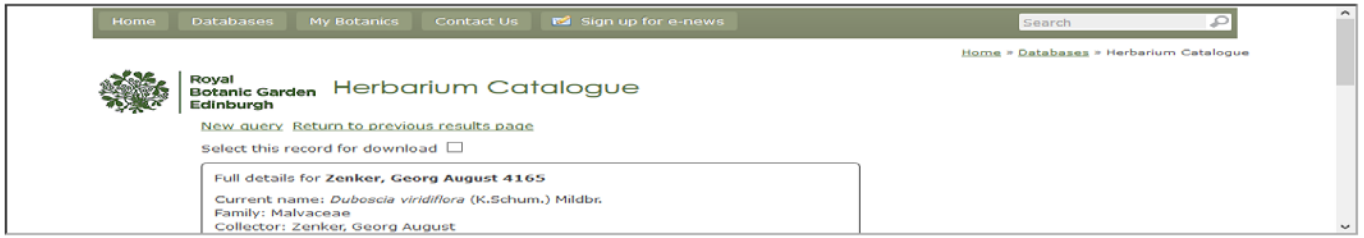
http://data.rbge.org.uk/herb/E00421509 [Check Now]

Checking URI format

- ✓ The scheme 'http' is correct
- ✓ The host has the domain name of 'data.rbge.org.uk'
- ✓ The path component is '/herb/E00421509'
- ✓ The URI lacks a query string component which is a good thing.
- ✓ Format of URI appears OK. Continuing test.

Requesting HTML Format Data

- ℹ Requesting HTML by passing 'Accept: text/html' header.
- ✓ Recieved 303 Redirect HTTP code.
- ✓ Redirect to URI: http://elmer.rbge.org.uk/bgbase/vherb/bgbasevherb.php?cfg=bgbase/vherb/fulldetails.cfg&specimens_specimen___num=469051



- ℹ Had 303 redirect when asking for HTML so will request RDF format.

Requesting RDF Format Data

- ℹ Requesting RDF by passing 'Accept: application/rdf+xml' header.
- ✓ Recieved 303 Redirect HTTP code.
- ✓ Redirect to URI: http://data.rbge.org.uk/service/rdf/herb.php?guid=http://data.rbge.org.uk/herb/E00421509
- ℹ Had 303 redirect so will request RDF data for parsing.

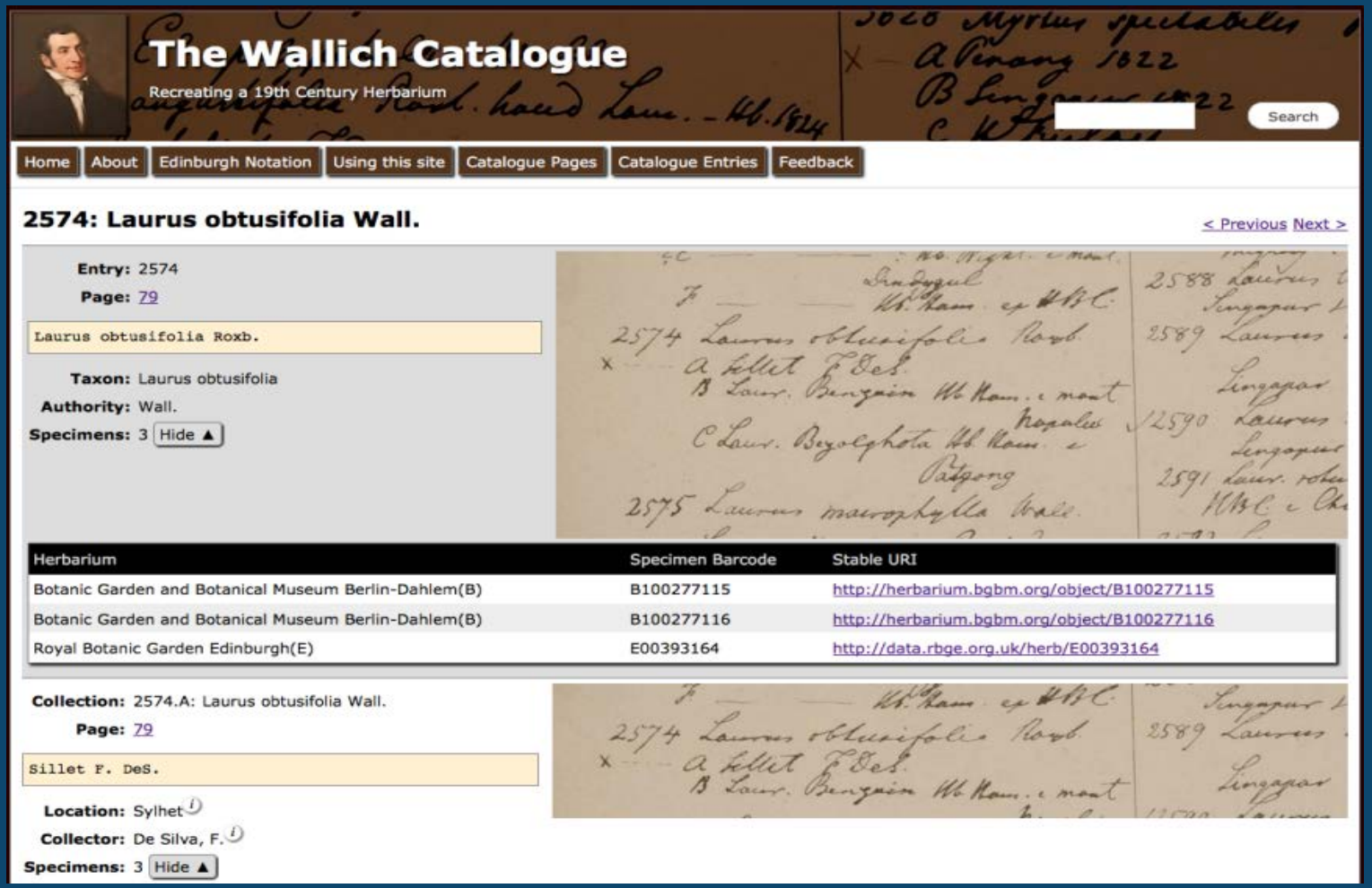
Parsing RDF

CETAF Specimen Preview Profile

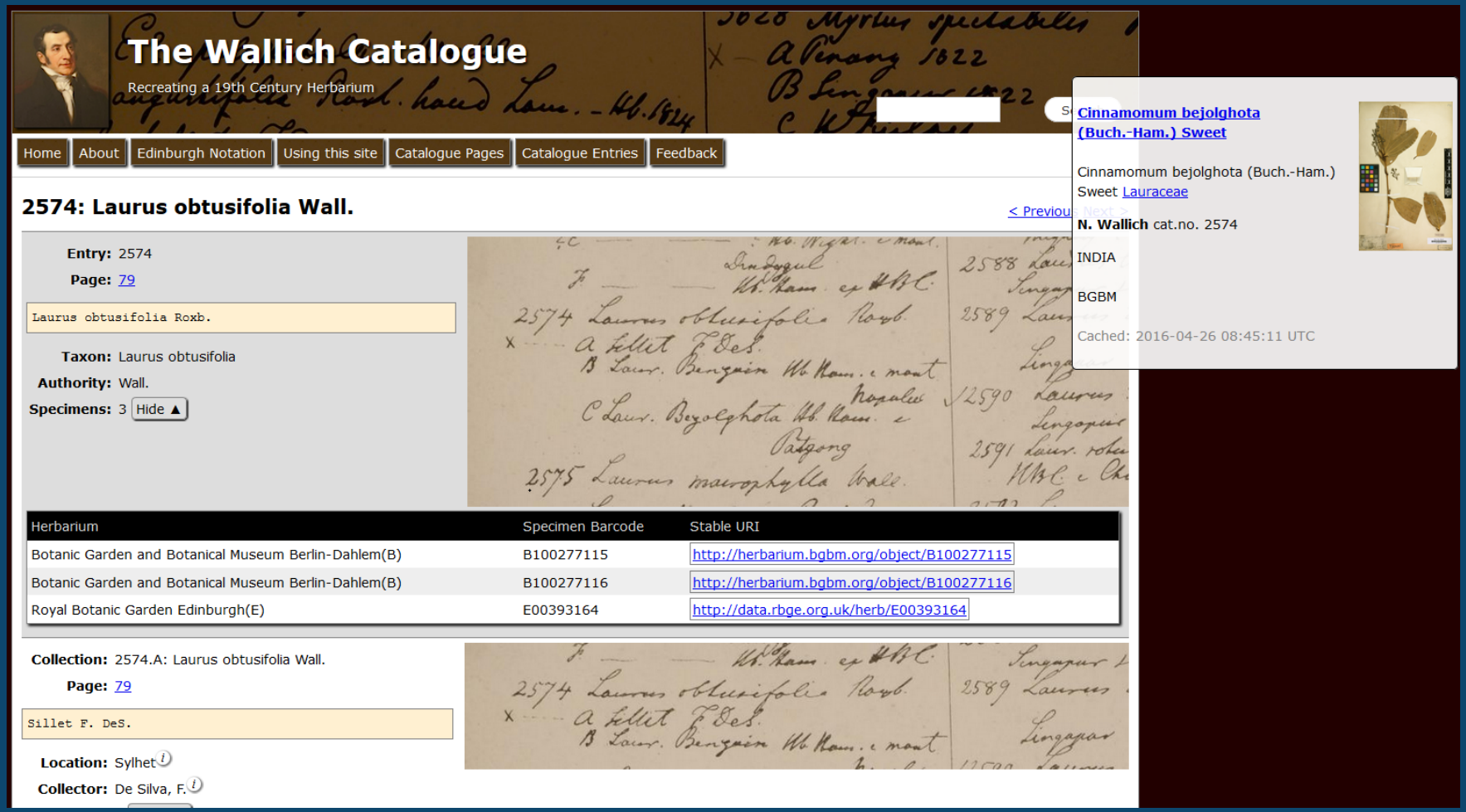
Prefered URI's are listed first followed by any depricated URIs that may contain the required values.

CSPP Element	Mandatory	RDF resource URI	Resource Expected	Value
title	Yes	http://purl.org/dc/terms/title	False	Zenker, Georg August #4165 Duboscia viridiflora (K.Schum.) Mildbr.
Kind of Material	No	http://purl.org/dc/terms/type	True	Specimen

Identifiers in use – pilot applications



Identifiers in use – pilot applications



Many Thanks!

Anton Güntsch

Freie Universität Berlin

Botanic Garden and
Botanical Musem Berlin

a.guentsch@bgbm.org

Freie Universität Berlin

BG BM | Botanischer Garten & Botanisches Museum Berlin