Globale Netze der Kooperation bei der Sacherschließung im naturkundlichen Bereich

Anton Güntsch, Botanischer Garten und Botanisches Museum Berlin-Dahlem

Zusammenfassung

Naturkundliche Sammlungen sind über Jahrhunderte mit großem Aufwand aufgebaut und bewahrt worden. Sie bilden das Rückgrad der systematischen Forschung und dienen als Datengrundlage für die Biodiversitätsforschung insgesamt. Mit der Entwicklung des BioCASE-Netzwerks ist es erstmalig in Europa gelungen, ein voll funktionsfähiges fach- und länderübergreifendes Netzwerk für Sammlungsinformation zur Verfügung zu stellen. Über BioCASE sind gegenwärtig Informationen zu 15000 Sammlungen und mehr als 5 Millionen Sammlungsobjekten verfügbar. Der BioCASE Datenservice ist gleichzeitig Teil der „Global Biodiversity Information Facility“ (GBIF), dem im Aufbau befindlichen weltweiten Netzwerk für Biodiversitätsdaten, über das bis heute über 70 Millionen biologische Sammlungsobjekte zugänglich gemacht wurden.

Hintergrund

Die in naturkundlichen Sammlungen weltweit vorhandenen biologische Sammlungsobjekte werden auf 2,5 Milliarden geschätzt (Duckworth & al. 1993), nur ein Bruchteil hiervon ist bisher elektronisch erfasst, und von diesen ist wiederum nur ein Teil über elektronische Netzwerke verfügbar gemacht worden. Wissenschaftlichen Arbeiten über taxonomische Gruppen geht daher häufig eine aufwändige Recherche über das Vorhandensein von Belegen in Sammlungen voraus, die dann ausgeliehen oder vor Ort in Augenschein genommen werden müssen. Mit der Digitalisierung und Vernetzung von Beleginformationen lässt sich der Rechercheaufwand wesentlich reduzieren, und hochauflösende digitale Belege können in vielen Fällen die Analyse des physischen Objekts erübrigen. Darüber hinaus eröffnen sich durch die Sammlungsvernetzung eine Vielzahl neuer Forschungsbereiche, die erst durch die schnelle Verfügbarkeit großer Mengen von Beleginformationen sinnvolle Ergebnisse liefern können. So wird zum Beispiel die Vorhersage der Verbreitung bestimmter Arten basierend auf Funden aus naturhistorischen Sammlungen verknüpft mit historischen Klimadaten erst aussagekräftig, wenn eine möglichst hohe Anzahl von Sammlungsinformationen für die untersuchten Arten über Netzwerke abgefragt und gemeinsam analysiert werden können.

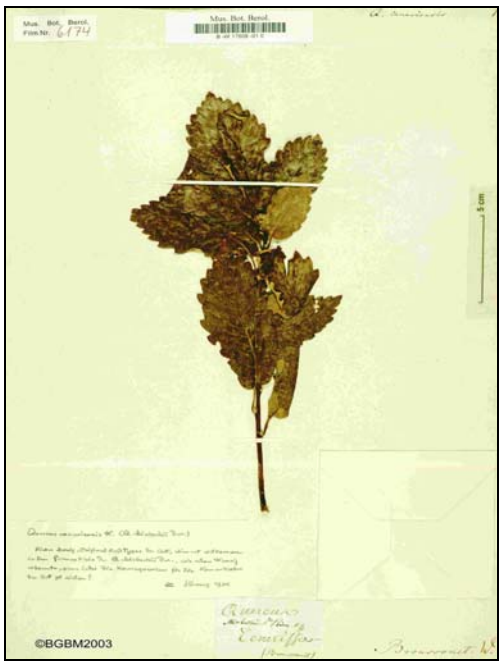

Abb.1: Digitalisierter Herbariumsbeleg, Copyright © 2005 BGBM

Mit der zunehmenden Verfügbarkeit von Internet- und Datenbanktechnologie in den Naturhistorischen Sammlungen Europas wurde der Plan, ein länder– und fachübergreifendes europäisches Sammlungsnetzwerk aufzubauen in den letzten 10 Jahren immer realistischer. Nachdem sich die Projekte CDEFD[1] (Berendsohn 1997) und BioCISE[2] (Berendsohn & Güntsch 2001) vorwiegend mit den Grundlagen für den Aufbau des Netzwerks befassten und unter anderem ein umfassendes Datenmodell für Biologische Sammlungen entwickelten (Berendsohn et. al. 1999), gelang es schließlich mit dem European Natural History Specimen Information Network ENHSIN[3] erstmals, einen Prototyp für ein gemeinsames Zugangssystem für verteilte und heterogene Sammlungsdatenquellen in vier europäischen Ländern zu realisieren (Güntsch 2003).

Die volle Implementierung des Netzwerks wurde schließlich im Rahmen des EU-Projekts BioCASE[4] (Berendsohn 2002) mit 25 Partnerinstitutionen aus 30 Ländern Europas und Israel realisiert. Dabei wurden die Knoten des Netzwerks so konzipiert und implementiert, dass Sammlungsnetzwerke mit beliebiger geographischer und fachlicher Ausprägung darauf zugreifen können.

Erfassung und Standardisierung vom Metainformation

Neben der zentralen BioCASE-Funktion, Informationen, die mit Sammlungsobjekten verknüpft sind, über das Internet verfügbar zu machen, sollte zusätzlich ein möglichst vollständiges Verzeichnis europäischer Sammlungen erstellt werden. Bereits das Vorgänger-Projekt BioCISE hatte mit Hilfe einer zentralisierten Umfrage rund 2800

[1] A Common Datastructure for European Floristic Databases, http://www.bgbm.org/CDEFD/default.htm
[2] Resource Identification for a Biological Collection Information Service in Europe, http://www.bgbm.org/biocise/
[3] http://www.bgbm.org/BioDivInf/projects/ENHSIN/default.htm
[4] Biological Collection Access Service for Europe, http://www.biocase.org

Sammlungen katalogisiert (Hahn 2000). Die Informationen wurden in einer Datenbank erfasst und im World Wide Web dynamisch publiziert[5] (Güntsch 2000).

Im Rahmen des BioCASE-Projekts wurde dieser Datenbestand dann durch ein neues System aus Nationalen BioCASE-Knoten weiter aufgebaut. Jeder nationale Knoten unterhält hierfür eine lokale Datenbank, die häufig bereits im Rahmen von nationalen Projekten aufgebaut worden war, in der die jeweiligen nationalen Metadaten gepflegt werden. Alle 24 Stunden werden die Daten mit dem zentralen BioCASE-Server, der am Botanischen Garten und Botanischen Museum Berlin-Dahlem unterhalten wird, synchronisiert, so dass dort immer ein aktuelles Verzeichnis sämtlicher Sammlungen vorgehalten wird.

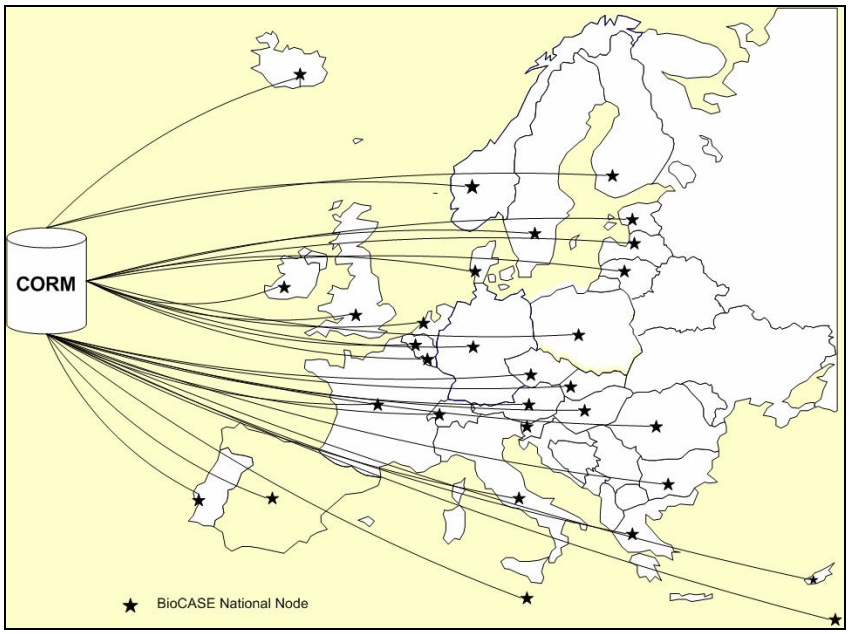

Abb2: Das BioCASE Netzwerk für Meta-Information

Um eine für den Import in die zentrale Datenbank notwendige Vergleichbarkeit der in den nationalen Knoten vorgehaltenen Daten zu erreichen, wurde ein XML-Schema entwickelt, das ein gemeinsames europäisches Format für Sammlungsmetadaten spezifiziert[6]. Dieses so genannte „Collection Profile“ stellt sicher, dass Metdaten, die im Netzwerk ausgetauscht werden, einheitlich strukturiert sind und dass die Bedeutung von Begriffen, die zur Beschreibung von Sammlungen verwendet werden, definiert ist.

Das BioCASE Collection Profile deckt Sammlungen, Organisationen und Netzwerke ab. Es liefert kontrollierte Vokabulare für XML-Elemente, die Gegenstand von knotenübergreifenden Abfragen sein können, und erlaubt die Darstellung von mehrsprachigen Inhalten. Der folgende Schema-Ausschnitt zeigt die Definition eines kontrollierten Vokabulars für Konservierungsmethoden.

[5] http://www.bgbm.org/biocise/DataBase/default.htm
[6] http://www.biocase.org/Doc/Results/results.shtml

```
<xs:simpleType name="PreservationMethodClassVoc">
        <xs:annotation>
            <xs:documentation>No treatment / Alcohol / Deep frozen / Dried / Dried and pressed /
                Formalin / Freeze-dried / Glycerin / Gum Arabic / Microscopic preparation / Mounted taxidermy /
                Pinned / Other
            </xs:documentation>
        </xs:annotation>
        <xs:restriction base="xs:NMTOKEN">
            <xs:enumeration value="no_treatment"/>
            <xs:enumeration value="alcohol"/>
            <xs:enumeration value="deep-frozen"/>
            <xs:enumeration value="dried"/>
            <xs:enumeration value="dried_and_pressed"/>
            <xs:enumeration value="formalin"/>
            <xs:enumeration value="refrigerated"/>
            <xs:enumeration value="freeze-dried"/>
            <xs:enumeration value="glycerin"/>
            <xs:enumeration value="gum_arabic"/>
            <xs:enumeration value="microscopic_preparation"/>
            <xs:enumeration value="mounted"/>
            <xs:enumeration value="pinned"/>
            <xs:enumeration value="other"/>
        </xs:restriction>
    </xs:simpleType>
```

Mit Hilfe der dezentralen Datenerfassung ist es gelungen, Metainformationen zu 4 Netzwerken, 2000 Institutionen und 15000 Sammlungen in Europa zu registrieren und aktuell zu halten. Die Daten sind unter anderem über das BioCASE Portal[7] abfragbar, dienen aber auch Klassifizierung und Registrierung von Sammlungen, die bereits Zugriff auf die individuellen Objekt-Informationen in ihren Sammlungsdatenbanken erlauben.

Beschreibung von Sammlungsobjekten mit ABCD

Obgleich das Netzwerk für Metainformation bereits eine wertvolle Informationsquelle für Sammlungsnutzer darstellt, so ist das eigentliche Ziel der BioCASE Initiative, die Verfügbarkeit von Sammlungsinformation auf Objektebene herzustellen. Auch hier macht die heterogene und verteilte Struktur existierender Sammlungsdatenbanken die Definition eines einheitlichen Datenaustauschformats nötig.

„Access to Biological Collection Data"[8] (ABCD) ist ein XML-Schema, das im Rahmen von BioCASE als Initiative der Taxonomic Databasing Working Group TDWG[9] und

[7] http://search.biocase.org/gui/
[8] http://www.bgbm.org/tdwg/CODATA/Schema/default.htm
[9] http://www.tdwg.org

CODATA[10] entwickelt und zum Austausch von Sammlungsdaten für beliebige biologische Sammlungstypen konzipiert wurde. ABCD beschränkt sich dabei nicht auf wenige Datenelemente, von denen man annimmt, dass sie in möglichst vielen Sammlungstypen relevant sind, sondern es wird versucht, mit einigen hundert Elementen verschiedenartigste Sammlungstypen umfassend abdecken zu können. Um Sammlungsdatenhaltern, deren Daten bisher noch wenig atomisiert sind, den Anschluss an Netzwerke zu erleichtern, wurden in vielen Bereichen zusätzliche Elemente eingeführt, mit denen sich Inhalte auch als Freitext übertragen lassen. Dieses Prinzip der „variablen Atomisierung“ hat in den ersten Jahren zu einem schnellen Wachstum des BioCASE Netzwerks geführt, da auf diesem Wege auch einfach strukturierte Datenquellen mit geringem Aufwand angeschlossen werden konnten.

Auf die Verwendung kontrollierter Vokabulare im ABCD Schema wurde in vielen Fällen bewusst verzichtet. Auf diese Weise sollte vermieden werden, dass kleinere Fehler in der verwendetet Terminologie, wie zum Beispiel Schreibfehler in Ländernamen, dazu führen, dass ganze Antwortdokumente invalide und von Portalen und Applikationen nicht verarbeitet werden. Stattdessen wurde für jedes ABCD-Element eine detaillierte und strukturierte Annotation eingeführt, welche dessen Semantik in einer menschen- und maschinenlesbaren Weise definiert. Diese Annotationen können neben einer Freitext-Beschreibung des Elements unter anderem Verweise auf Elemente gleicher Bedeutung in anderen Standards, Beispiel-Einträge, Regeln für die Erzeugung korrekter Einträge und kontrollierte Vokabulare enthalten. Der folgende Schemaausschnitt zeigt die Definition des Elementes *AreaName* in der Version 2.0 des ABCD-Schemas:

```
<xs:element name="AreaName" type="StringL" minOccurs="0">
    <xs:annotation>
        <xs:documentation>Name of a geographic, geomorphological, geoecological, or administrative area.
        </xs:documentation>
        <xs:appinfo>
            <sea:FullName>Area name</sea:FullName>
            <sea:Reviewer>Berendsohn 02 Mar 2003</sea:Reviewer>
            <sea:ExistingStandard>Darwin Core 2: Locality.</sea:ExistingStandard>
            <sea:ExistingStandard>Darwin Core 2: State Province.</sea:ExistingStandard>
            <sea:ExistingStandard>Darwin Core 2: County.</sea:ExistingStandard>
            <sea:Example>Atlantic coast</sea:Example>
            <sea:Example>Santa Cruz</sea:Example>
            <sea:Example>Weinheim</sea:Example>
            <sea:Example>Andes</sea:Example>
            <sea:Example>El Imposible</sea:Example>
        </xs:appinfo>
    </xs:annotation>
</xs:element>
```

[10] http://www.codata.org

Neben seiner Rolle als zentrales Datenaustauschformat im BioCASE Netzwerk ist ABCD auch ein Standard-Schema für das Sammlungsdaten-Netzwerk der Global Biodiversity Information Facility. In diesen Netzwerken wird gegenwärtig die Version 1.2 des Schemas verwandt, die Version 2.0 steht kurz vor der Verabschiedung als TDWG-Standard im September 2005.

Verknüpfung heterogener Sammlungsdatenbanken

Für die Vernetzung heterogener Sammlungsdatenbanken ist neben der Spezifikation eines gemeinsamen Vokabulars, das mit dem ABCD Schema definiert wurde, eine Festlegung einer einheitlichen Anfragesprache, die unabhängig von den nativen Anfragesprachen der angeschlossenen Datenbanken ist, notwendig. Das BioCASE-Protokoll spezifiziert eine solche Sprache, die wie das ABCD Datenschema XML verwendet. Es werden lediglich drei mögliche Anfrage-Typen festgelegt:

- **Capabilities** erfragt vom Datenanbieter, für welches Datenschema er konfiguriert wurde und welche Datenschema-Elemente er liefern kann.
- **Scan** erfragt für ein gegebenes Datenelement, welche distinkten Werte hierfür in der Datenbank vorhanden sind. Mit dieser Abfrage lässt sich also zum Beispiel für ein Element, das Länderbezeichnungen enthält, die Liste aller verschiedenen Länderbezeichnungen in dem entsprechenden Feld der Sammlungsdatenbank erzeugen.
- **Search** ist der eigentliche Suchbefehl, mit dem Sammlungsdaten, die dem übergebenen Suchmuster entsprechen, in der angeschlossenen Datenbank gefunden werden können.

Abbildung 3 zeigt vereinfacht den Ablauf einer BioCASE-Abfrage. Eine mit einer Nutzerschnittstelle im World Wide Web formulierte Suchanfrage wird von einem Portal als BioCASE-konforme Suchanfrage an die europaweit verteilten Datenanbieter weitergeleitet. Jeder dieser Datenanbieter hat ein Softwaremodul installiert, das die BioCASE-Anfrage in die native Anfragesprache der jeweiligen Datenbank und Konzepte der Datenbank in ABCD-konforme Konzepte übersetzt. Die Antwortdaten werden dann als ABCD-Dokument an das Portal zurückgesendet, und das Ergebnis wird in lesbarer Form in der Nutzerschnittstelle dargestellt.

Alle Software-Komponenten sind frei verfügbar und können auch unabhängig vom BioCASE Projekt für Netzwerke ganz anderer Art außerhalb der Biologie eingesetzt werden (Döring & Güntsch 2003).

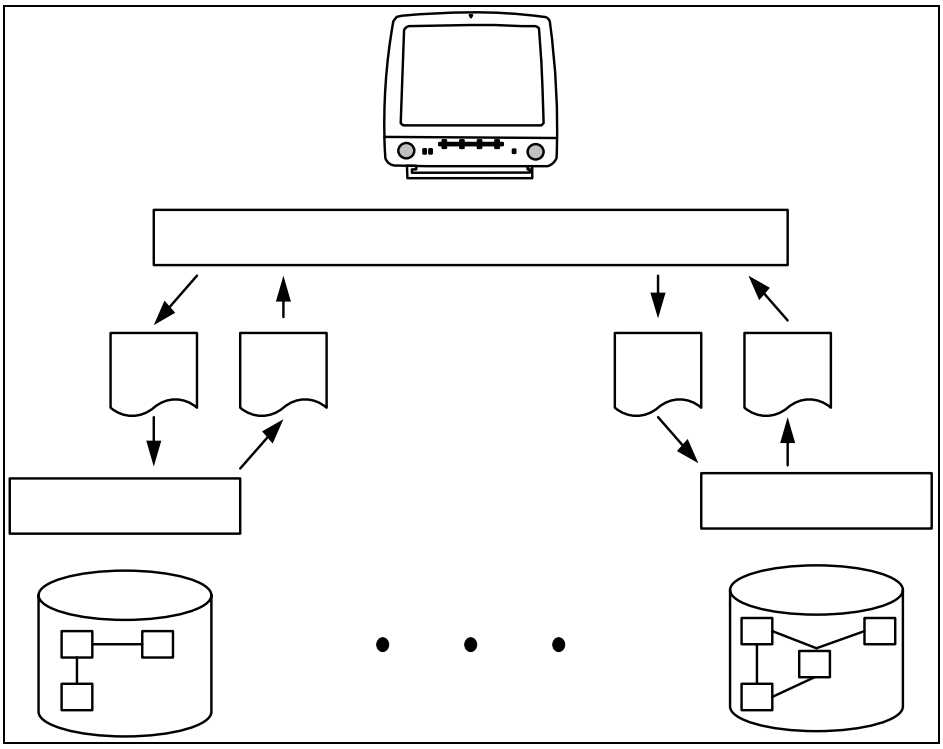

Abb 3: Vereinfachter BioCASE Datenfluss

Das BioCASE Protokoll und die zugehörigen Softwarekomponenten werden inzwischen in mehreren europäischen Netzwerken eingesetzt. Prominentester Nutzer ist allerdings das Datenportal der Global Biodiversity Information Facility, mit dem von BioCASE-Knoten gelieferte ABCD-Daten abgefragt werden können (s. Abb. 4).

Das Portal realisiert zusätzlich die Integration der Sammlungsdaten mit den in GBIF vorgehaltenen taxonomischen Daten, so dass es möglich ist, durch taxonomische Bäume zu der gewünschten Information zu navigieren. Zusätzlich wurde ein Modul zur Verfügung gestellt, mit dem man aus ABCD-Sammlungsdaten dynamisch Landkarten mit Markierungen der jeweiligen Fundorte erzeugen lassen kann.

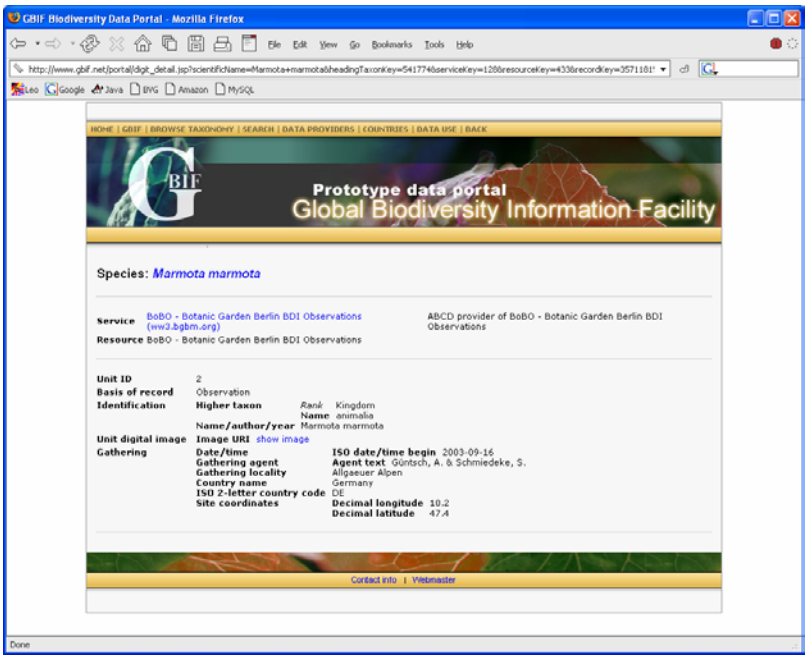

Abb 4: Zugriff auf BioCASE-Daten über das GBIF-Portal

Ausblick

Mit dem BioCASE Projekt ist es erstmals gelungen, Millionen von Objekten aus europäischen biologischen Sammlungen einheitlich im Internet zugänglich zu machen. Folgeprojekte werden diese Datenbasis mit der zunehmenden Digitalisierung von Sammlungsbelegen und dem Anschluss neuer Datenanbieter weiter verbreitern und die Implementierung von völlig neuen Anwendungen ermöglichen. Dabei wird es darauf ankommen, aufbauend auf die existierenden Portalprototypen, spezialisierte Portale zu entwickeln, die Sammlungsdaten mit existierendem Fachwissen in geeigneter Weise verknüpfen. Dies können zum Beispiel taxonomische Thesauri sein, mit deren Hilfe Anfragen durch den Einsatz von Synonym-Listen erweitert werden, geographische Informationssysteme (GIS) können dazu genutzt werden, Sammlungsreisen auf der Grundlage von Sammlungsdaten nachzuvollziehen und fehlende Geo-Referenzen nachzutragen. Annotations-Module werden Wissenschaftler in die Lage versetzen, Ergebnisse ihrer Arbeit elektronisch zu übermitteln, so dass sie zusammen mit dem Beleg elektronisch verfügbar gemacht werden können.

Literatur

Berendsohn, W. G. 1997: A Common Datastructure for European Floristic Databases (CDEFD). Pp. 251-253 in: Hoeveler, A. & Cresti, M. (ed.), Biotechnology (1992-1994), Final Report Vol. 1. Luxemburg.

Berendsohn, W.G., Anagnostopoulos, A. Hagedorn, G., Jakupovic, J., Nimis, P.L., Valdés, B., Güntsch, A., Pankhurst, R.J. & White, R.J. 1999: A comprehensive reference model for biological collections. Taxon 48: 511-562.

Berendsohn, W. G. & Güntsch, A.. 2001: Resource Identification for a Biological Collection Information Service in Europe (BioCISE). In: Proceedings of the IX OPTIMA Meeting Paris, May 1998, BOCCONEA 13.

Berendsohn, W. G. 2002: BioCASE - A Biological Collection Access Service for Europe. Alliance News 29(6): 6-7.

Döring, M. & Güntsch, A. 2003: Technical introduction to the BioCASE software modules. 19th annual meeting of the Taxonomic Databases Working Group (TDWG 2003), Abstract, Oeiras.

Duckworth, W. D., Genoways, H. H. Rose, C. L. 1993: Preserving natural science collections: Chronicle of our environmental heritage. – Washington.

Güntsch, A. 2000: The BioCISE collection catalogue. In: Berendsohn, W.G. (ed.): Resource Identification for a Biological Collection Information Service in Europe, Botanic Garden and Botanical Museum Berlin-Dahlem.

Güntsch, A. 2003: The ENHSIN Pilot Network. In: Scoble, M. (ed): ENHSIN - The European Natural History Specimen Information Network. The Natural History Museum, London.

Hahn, A. 2000: Information Resources – The BioCISE Survey. In: Berendsohn, W.G. (ed.): Resource Identification for a Biological Collection Information Service in Europe, Botanic Garden and Botanical Museum Berlin Dahlem.

Dr. Anton Güntsch
Botanischer Garten und Botanisches Museum Berlin-Dahlem
Abt. Biodiversitätsinformatik und Labore
Königin-Luise-Str. 6-8, 14191 Berlin
http://www.bgbm.org
a.guentsch@bgbm.org