



From Static Corpora to Dynamic Collections



Jacqueline Ringersma

Peter Wittenburg

MPI for Psycholinguistics

DOBES Archive

(DOKumentation BEdrohter Sprachen)

(Documentation of Endangered Languages)

(funded by the VolkswagenFoundation)



Background



- some time ago we were informed by the OLAC service provider that they will not harvest our 40.000 metadata descriptions and that they only would harvest records representing “corpora”
 - to behave nicely we created for the finishing DoBeS documentation teams OLAC compliant descriptions
 - so instead of 40.000 records we now offer 15 records or so
-



Background



Question:

- What can we do? Can we do more?

Principle question to address:

- What is a corpus – is it still a valid concept

The answer seems to be a fuzzy Yes (No/Yes),
but with a tendency towards a clearer: No



Classical Corpus Concept



classical corpus:

A **set of language resources** based on:

a careful **design**

to serve a certain (scientific) purpose

to answer a number of predefined (scientific) questions



Classical Corpus Concept



The design covers **different dimensions**:

- type of genres included etc
- type of language variety included
- type of analysis (annotations) included

- amount of resources along these dimensions to have a solid basis for statistics etc

Typically such a corpus was carefully designed and prepared in a first phase leaving further scientific analysis etc to a second phase

Traditionally the original recordings were not accessible anymore i.e. the transcriptions etc. were seen as the basis for further work



Many good Examples



• Dutch Spoken Corpus

The screenshot shows a hierarchical tree structure for the Dutch Spoken Corpus. The root node is 'CGN'. Under 'CGN', there are several sub-nodes: 'CGN Documentation in English', 'All', 'DVDs', 'Annotation types', 'Components', 'Regions', 'Speaker Sexes', and 'Speaker Ages'. The 'Annotation types' node is expanded, showing 'phonetic annotations', 'prosodic annotations', and 'syntactic annotations'. The 'Components' node is also expanded, showing a list of communication types such as 'spontaneous conversations (face-to-face)', 'interviews with teachers of Dutch', 'spontaneous telephone dialogues (recorded via a switchboard)', 'spontaneous telephone dialogues (recorded on MD with local recording)', 'simulated business negotiations', 'interviews/discussions/debates (broadcast)', '(political) discussions/debates/meetings (non-broadcast)', 'lessons recorded in a classroom', 'live (eg sport) commentaries (broadcast)', 'newsreports/reportages (broadcast)', 'news (broadcast)', 'commentaries/columns/reviews (broadcast)', 'ceremonious speeches/sermons', 'lectures/seminars', and 'read speech'. The 'Regions', 'Speaker Sexes', and 'Speaker Ages' nodes are also expanded, showing a list of regions, speaker sexes, and speaker ages respectively.

- Design of corpus on the basis of:**
- Annotation types
 - Communication type components
 - Region of origin of speaker
 - Speaker gender
 - Speaker age



Many good Examples



- Dutch Spoken Corpus
- Dutch Bilingualism Corpus
- Dutch Diachronic Corpus of the time X to Y
- British National Corpus
- Lund Corpus
- Brown Corpus
- Stoll Corpus
- ESF Second Learner Corpus
- Switchboard Corpus
- ...

In general such corpora were **checked on validity** and one could speak about **published versions**



Classical Corpus Concept



Traditional corpora:

Static – finite size bodies of text

A standard reference for the language concerned
(Machine readable)

OAIS – open archive international standard (ISO)

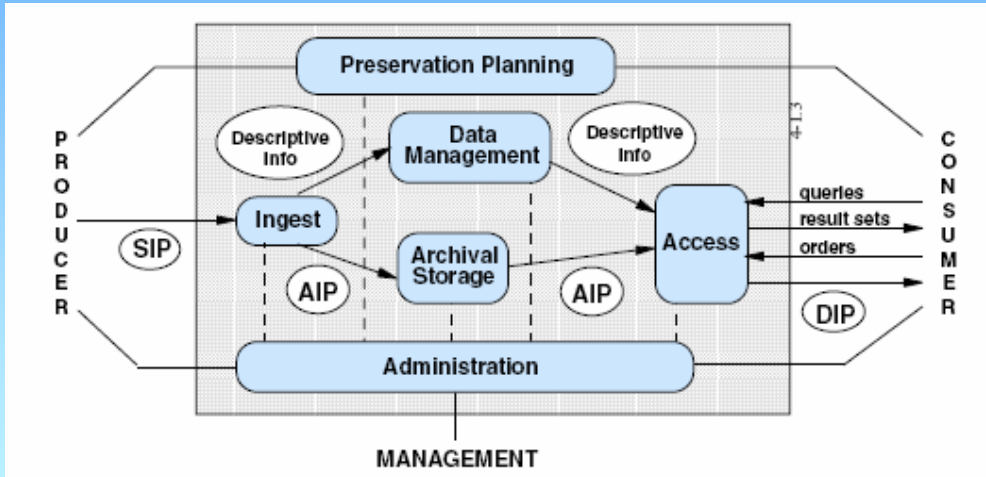


Classical Corpus Concept



OAIS – open archive international standard (ISO)

Model: 6 functional entities



Main information concept:

AIP - Archival Information Package



Classical Corpus Concept



OAIS – open archive international standard (ISO)

Archival Information Package

Archival Information Package (AIP): An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS.

So: defined as a static body (?)



Overcapacity and Extension



- What can we currently see in research institutes such as MPI, in Documentation programs and on the Web?

“Overcapacity” and Extension Phenomenon

too much data to be handled close in time

- making recordings, storing and accessing them is so easy
- people make many recordings of situations that seem to be interesting/relevant
(in E-documentation this makes also very much sense)

only part of the data is analyzed/annotated

- many will be sparsely annotated to meet the immediate needs

in documentation resources will be added after project end

- collections will dynamically grow in different dimensions
 - cannot speak about “finished/published” corpora
-

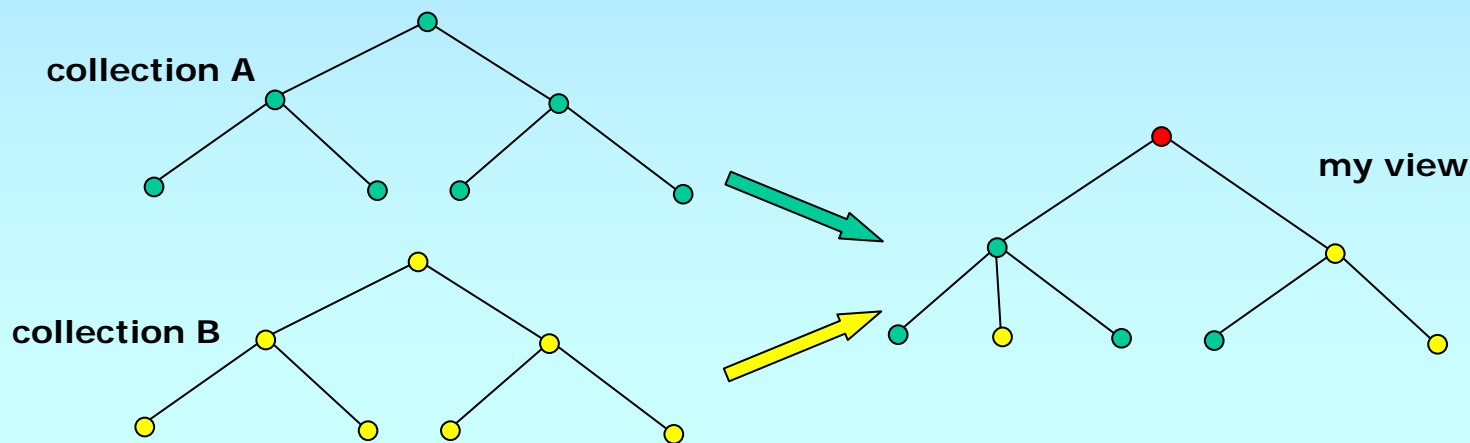


Different Views



View Phenomenon

- due to access to original recordings collections are rich
- cover much more information than “original intention”
- examples
- many sound recordings that were annotated “linguistically” can be subject of phonetic studies (prosody, tones, ...)
- many video recordings that were annotated “linguistically” can be subject of multimodality studies (gesture, facial expr, ...)
- many video recordings that were annotated “linguistically” can be subject of anthropological studies
- researchers create “personal” collections/views





Content Enrichment



- **Enrichment Phenomenon**

- resources are commented
 - people will draw relations between fragments of resources
 - this is in accordance with **Live Archives** idea (please support this: www.mpi.nl/!!)
 - move from linguistic documentation to knowledge documentation (see Gaby Cablitz and Bruce Birch)
 - active involvement of language community (extensions of all sorts to be expected)

 - MPI Team (for example) has two tools supporting these ideas:
 - **LEXUS** as anchor point for rich encyclopedia+linguistics
 - **ADDIT** for commentary and relation drawing

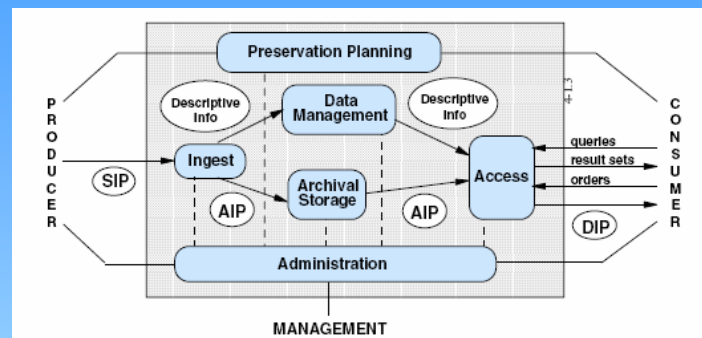
 - in particular relations (linguistically, anthropologically etc driven) create completely different views on collections
-



Towards a dynamic Corpus Concept



Basis still is the OAIS Model, but beyond:



giving access in the **Live** Archives sense

- archive as a center of ORGANIZED information
- archive as a center for continuous extension
- archive as a center of collaboration and interaction
- archive as a center for commentary and relation drawing (enrichment)
- archive as a center supporting cross-corpus/language work

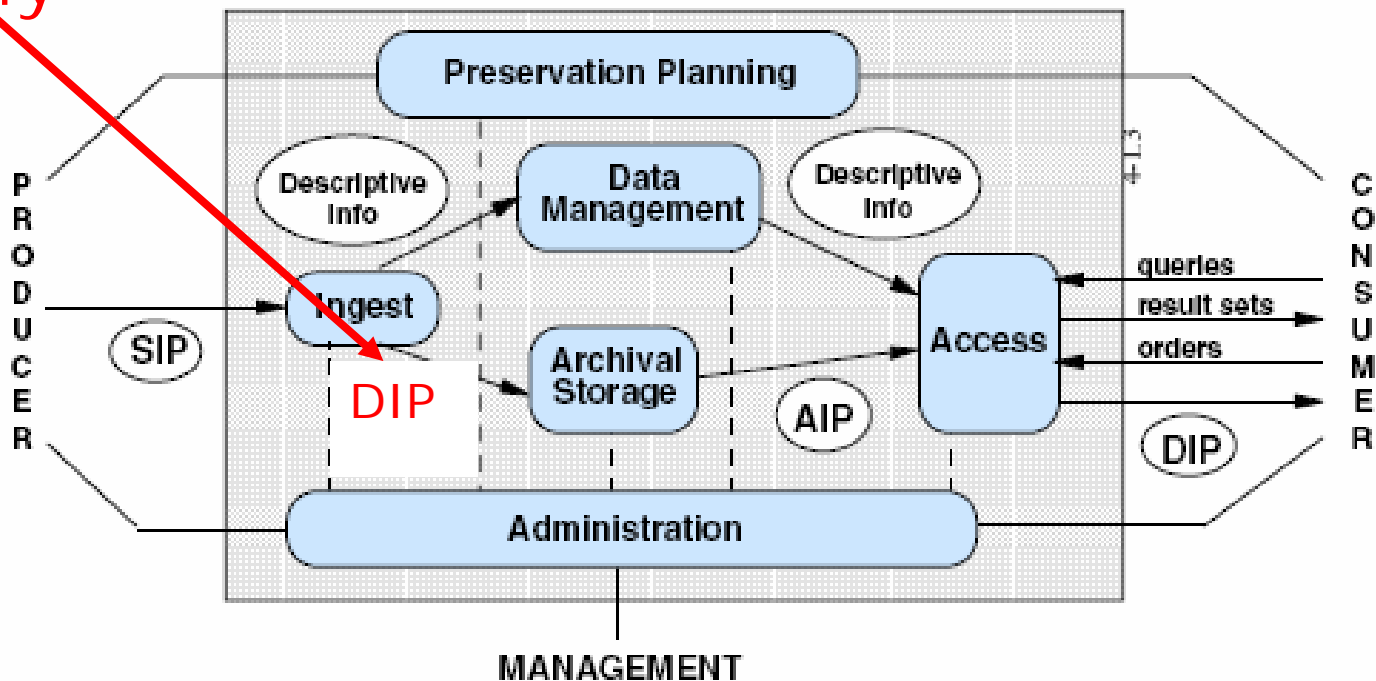


Towards a dynamic Corpus Concept



Basis still is the OAIS Model, but **interaction** is continuous

modify



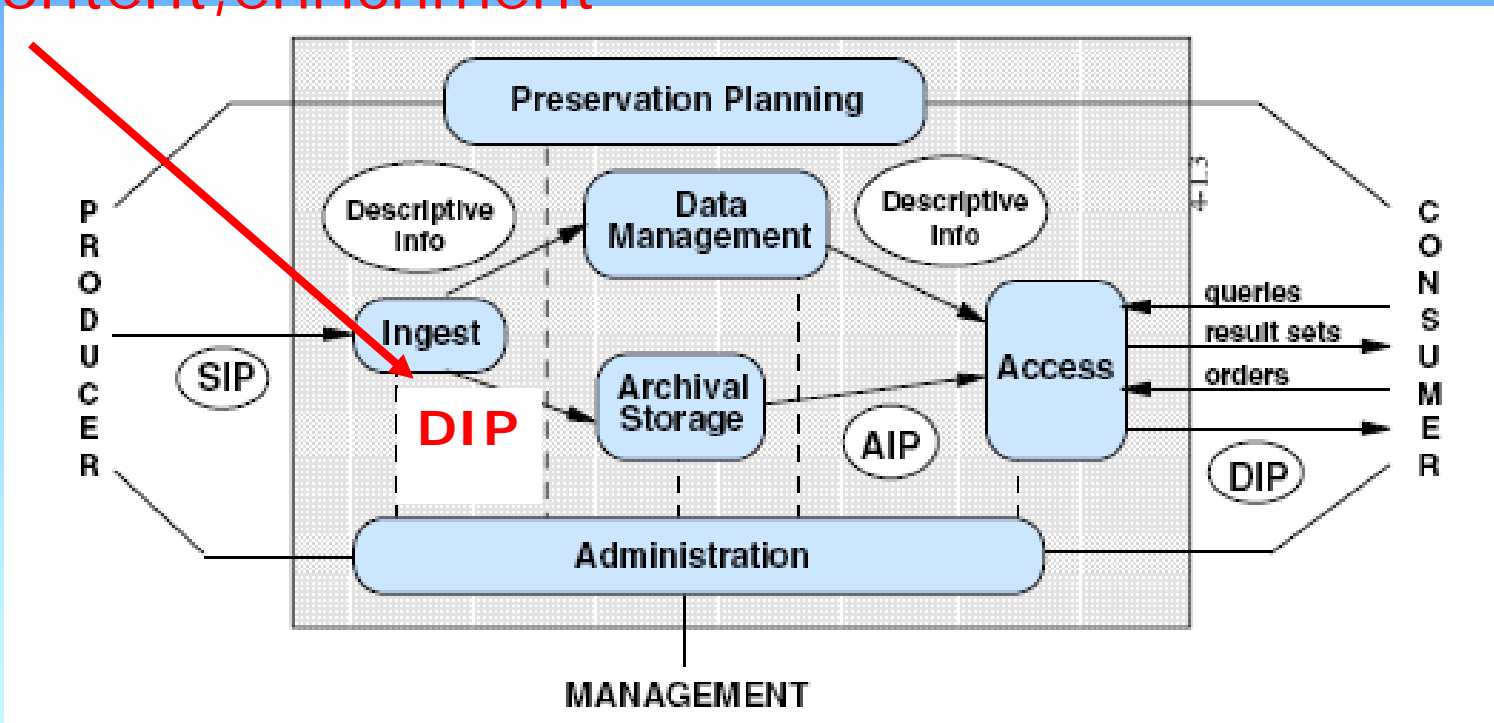


Towards a dynamic Corpus Concept



Basis still is the OAIS Model, but **interaction** is continuous

add content, enrichment





Summary



- the notion of “corpus” is fuzzy for most resources at the MPI
 - increasingly often we see **just different views on dynamic collections**
 - finally we will have both
 - in field linguistics and documentation dynamic collections
 - in some areas (language model parameter identification) people need evaluated and published corpora
 - crawling on the Web (ODIN etc) also will lead to arbitrary collections – can’t call the result a corpus
 - therefore bad answer to OLAC:
 - the researchers are responsible for resource bundling into collections
 - can’t create “corpus” nodes except for very obvious cases
-