Digital Library and Document Server

Abdel Belaïd

Digital Library and Docum Server

Calliope, Muse of eloquence and poe

The Calliope Project

A. Belaïd

The Framework

R&D Collaboration

- IMAG + INRIA + RXRC (Xerox)
- First collaboration success : Callimaque
  - 3000 thesis & technical reports : 1950 - 1993
  - RXRC : multilingual interrogation for anglophone readers
  - IMAG : 320 journals, annual budget 150 000 $

The problem

- INRIA Displacement, deprived researchers of doc access

The Solution

- Not duplicate the center
- Automatic system : access to IMAG document center

Constraints

nfrastructure :

- Investment : to minimize the cost in material
  - Use : Net infrastructure (connection flow 128 K bits/s)
  - Permit : heterogeneous client station (Macintosh, PC, Unix)
- Quality : to deliver articles with a quality similar to the photocopy
  - Postscript level 2

esearch : to take advantage of this opportunity

- To go beyond the paper model
  - To better apprehend the mass of scientific information
  - To integrate sources of heterogeneous documents
  - To open the research on data indexing, data mining, ...

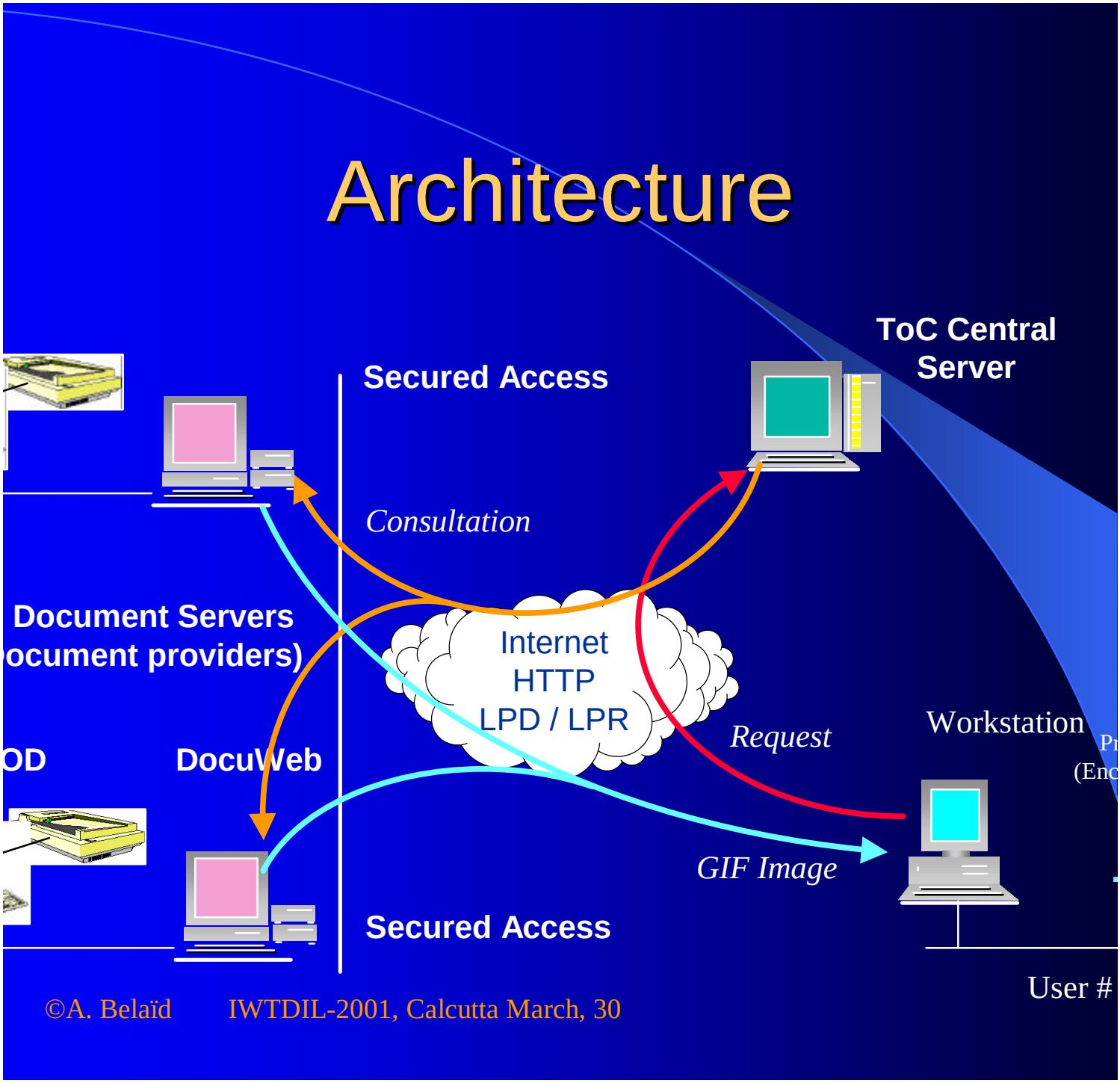
The Calliope service

The solution

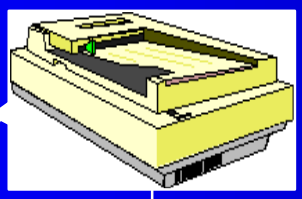
- To develop the concept of telephotocopy

Functionality

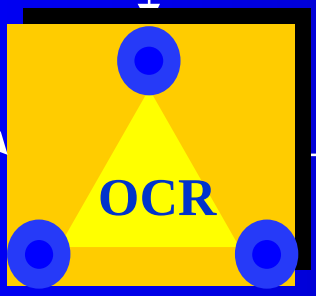
- Access to table of contents server
- Asking of telephotocopy by :
  - Simple click on the article
  - Textual Retrieval on paper titles, authors, journal title
- Services
  - Availability signaled by email
  - Get the paper in an image format
  - Printing at the request of part or total paper
  - Electronic subscription


Architecture
ToC Central Server
Secured Access
Consultation
Document Servers
ocument providers)
Internet
HTTP
LPD / LPR
Request
Workstation
OD
DocuWeb
GIF Image
Secured Access
User #
©A. Belaïd
IWTDIL-2001, Calcutta March, 30

Recognition

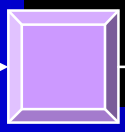


Scanner



OCR
Combination

XmlLayout



ToC
Analysis

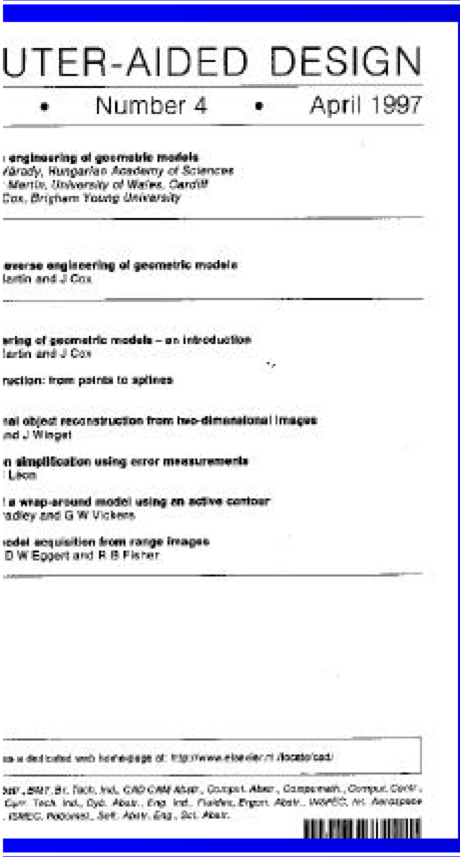
XmlToC

Correction

XmlCal

Calliope

ToC Examples


UTER-AIDED DESIGN
• Number 4 • April 1997

engineering of geometric models
Várady, Hungarian Academy of Sciences
Martin, University of Wales, Cardiff
Cox, Brigham Young University

everse engineering of geometric models
Martin and J Cox

ering of geometric models – an introduction
Martin and J Cox

ruction: from points to splines

al object reconstruction from two-dimensional images
nd J Winget

n simplification using error measurements
Leon

a wrap-around model using an active contour
adley and G W Vickers

odel acquisition from range images
D W Eggert and R B Fisher


Applicable Algebra in Engineering, Communication and Computing

A 21053

Vol. 8 No. 6 1997

Online edition in LINK Computer Science Online Library http://link.springer.de

AAECEW 8(6) 437–562 (1997)
December 1997
Printed on acid-free paper

LINK Now available online http://link.springer.de

0938-1279(199712)8:6;1-V

Springer

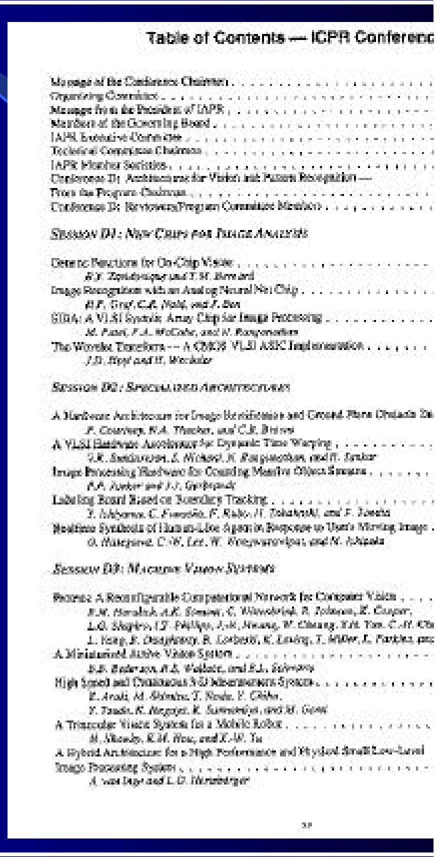
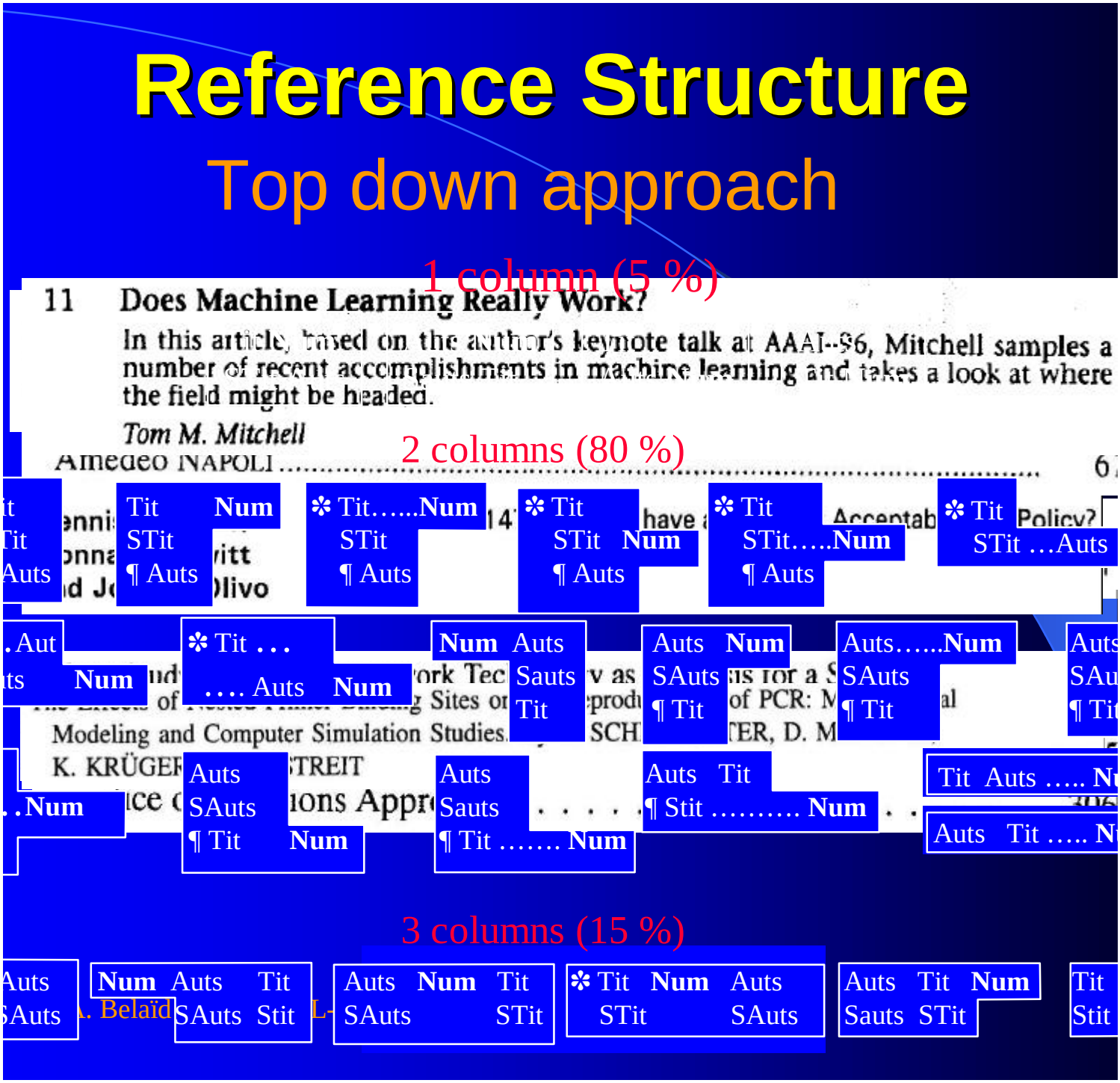


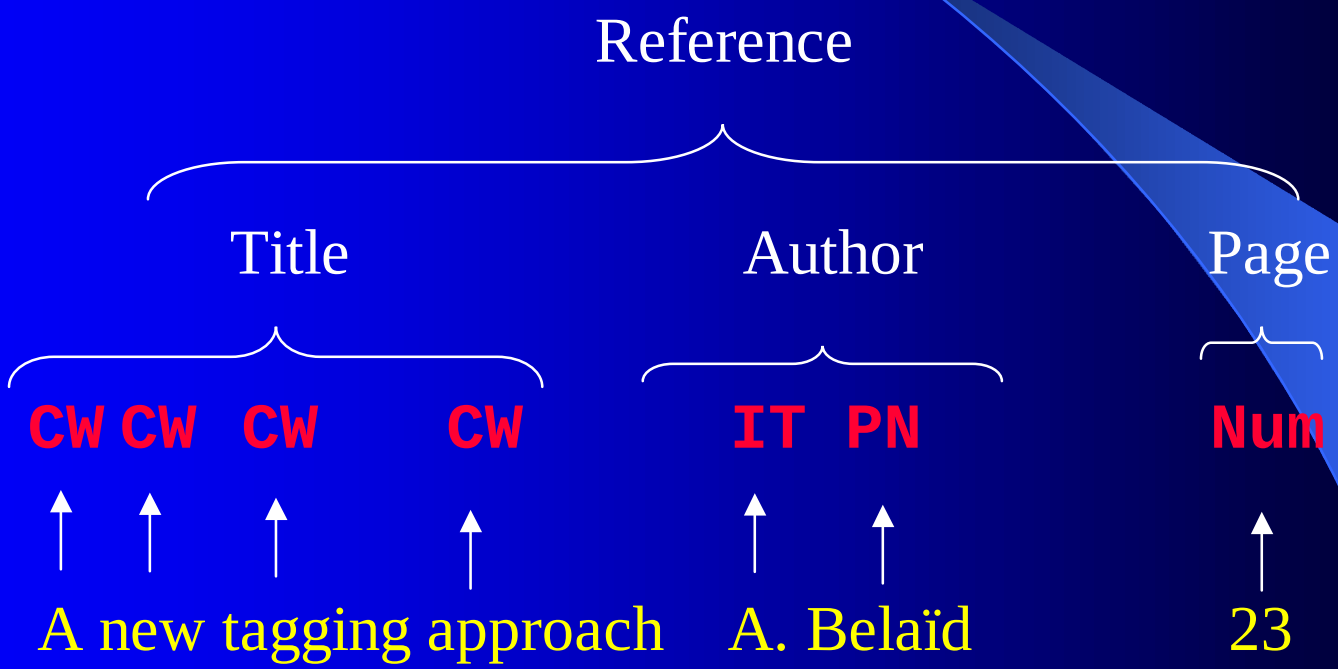
Table of Contents — ICPR Conference

Reference Structure

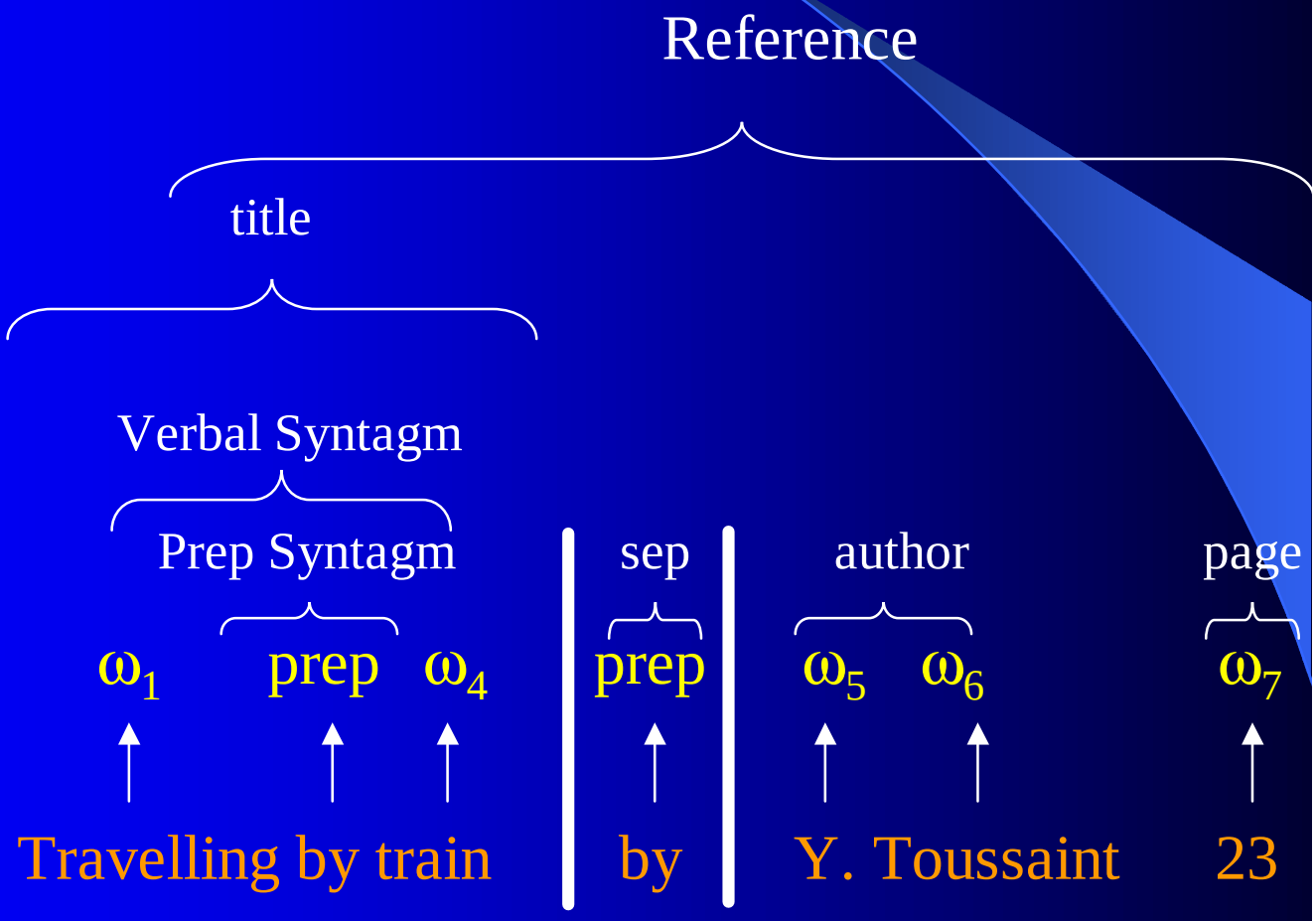
Top down approach



Idea : Bottom-Up Approach



Linguistic Help



Morpho-syntactical Labeling

Role :

- Annotate text to mark important elements of a corpus
- Associate these elements in more complex structures (terms) by syntactical agglutination

Initial application domain : terminology

- Search for terms in a document (comp of several words)

development/CN of/PREP systems/PCN

We can have :

Rule 1 : term ::= CN of/Prep PCN

Advantages

raining :

- Most of the systems work with a usable learning
- It is possible to enrich the system and to adapt it for its own use by doing an adequate learning

rediction :

- To have a system able to predict a label for an unknown term
- Example : Brill rule for French : **radical+ion Þ CN**

ocation Principle : labeling error correction by local rul

- **beginning of a sentence + capital Þ CN,**
- **capital in the middle of a sentence Þ PN**
- **JJ of/Prep CN Þ JJ = CN**

Application

l Representation for ation in Animats	J. Prescon
oring Strategies for dded Agents: Experiments nalysis	S. Ackin and R. Cohen

Prep p UN	UN . UN
N Prep N : PCN N	UN . UN CC UN. PN

Remarks

- Linguistic defects in lab. come from particulariti
  - Terms of the structure : not conform to text notion
  - Word structure : not conform to the morphology of th current language

```
S/UN ./. Ackin/UN
```

- Solutions :
  - Retrain the system: costly and not adapted
  - Choose a less fine grain for labeling and a set of adapted labels to the application

Adapted Labels

Physical Labels

Blank Line	BL	Beginning Line
Long Space	EL	Ending Line
Current Line	NL	Next Line
Numerical Line	PL	Previous Line
Dotted Line	PT	Punctuation
Tabulation	PU	Item

Morphological Labels

Common Noun	AR	Article
Proper Noun	FN	First name
Initial	CC	Connector
Preposition	UN	Unknown

Reference Location

Typographical aspect combination

ge number detection

Regularity

- Location : position, before/after TB/SL
- Alignment
- Continuity

eference beginning/ending determination

CL = BL : NS & (PL=SL | PL ' EL)

CL = EL : NS & (NL=SL | NL ' BL)

olumn identification

COL = {**CL** : ≅alignment, ≅position, >**TB**, <**TB**}

ARTICLES

Primary Labeling

			CN SL SL CN SL
presentation for in Animats	Tony J. Prescon	85	AJ CN PR TB FN IT PN TB NS CN PR UN SL
Strategies for Agents: Experiments sis	Marc S. Ackin and l 'aul R. Cohen	125	CN CN PR TB FN IT UN CC UN I CN CN PT CN CC CN SL
the competitors	Luc Steels	173	CN AR CN TB FN PN TB NS SL UN SL
Adaptations and ry Epistemology	James H. Fetzer	201	CN CN CC TB PN IT UN TB NS CN CN

Combination of linguistic aspects

orming rules :

- **Author(Aut): {IT ; FN ; PN}**
- **Title(Tit): {CW}-{negative cuttings} :DET,P**

egrouping rules :

- **Author : {Aut; CC ; BY ; PT}**
- **Title : {Tit ; CC ; PT}**

ontinuity rules :

- **Author : {IT ; FN} + EL + PN þ Aut + EL + A**

ative Method for Improved Protein Structural Motif Recognition.
BERGER and M. SINGE

cal Adaptations and Lee K. Wang and James H.
onary Epistemology Fetzer

Author Correction :

Re-establishment of the syntax

glutination rules:

ut + UN + Aut Þ Aut

ension rules:

ut + UN + {PN} + DL|EL Þ Aut + DL|EL

icipation rules:

Tit + EL + TB + UN+ Þ Tit + EL + TB + AutÞ

utÞ + CC + FN|IT + UN Þ Aut

Title Correction :

e-establishment of the syntag

glutination rules:

+ + UN + Tit+ Þ Tit+

+| Tit Þ + Prep + Aut + Tit+ Þ Tit+

Þ + Prep + Tit+ Þ Tit+

ension rules + anticipation :

+ Prep|Det + UN+ Þ Tit Þ

+ Prep|Det + Tit Þ Ü Tit

+ EL + TB + UN+ Þ Aut + EL + TB + Tit Þ

ing and evolulion in networks

es extracted from the article of /IN Claude Allègre « What I want »

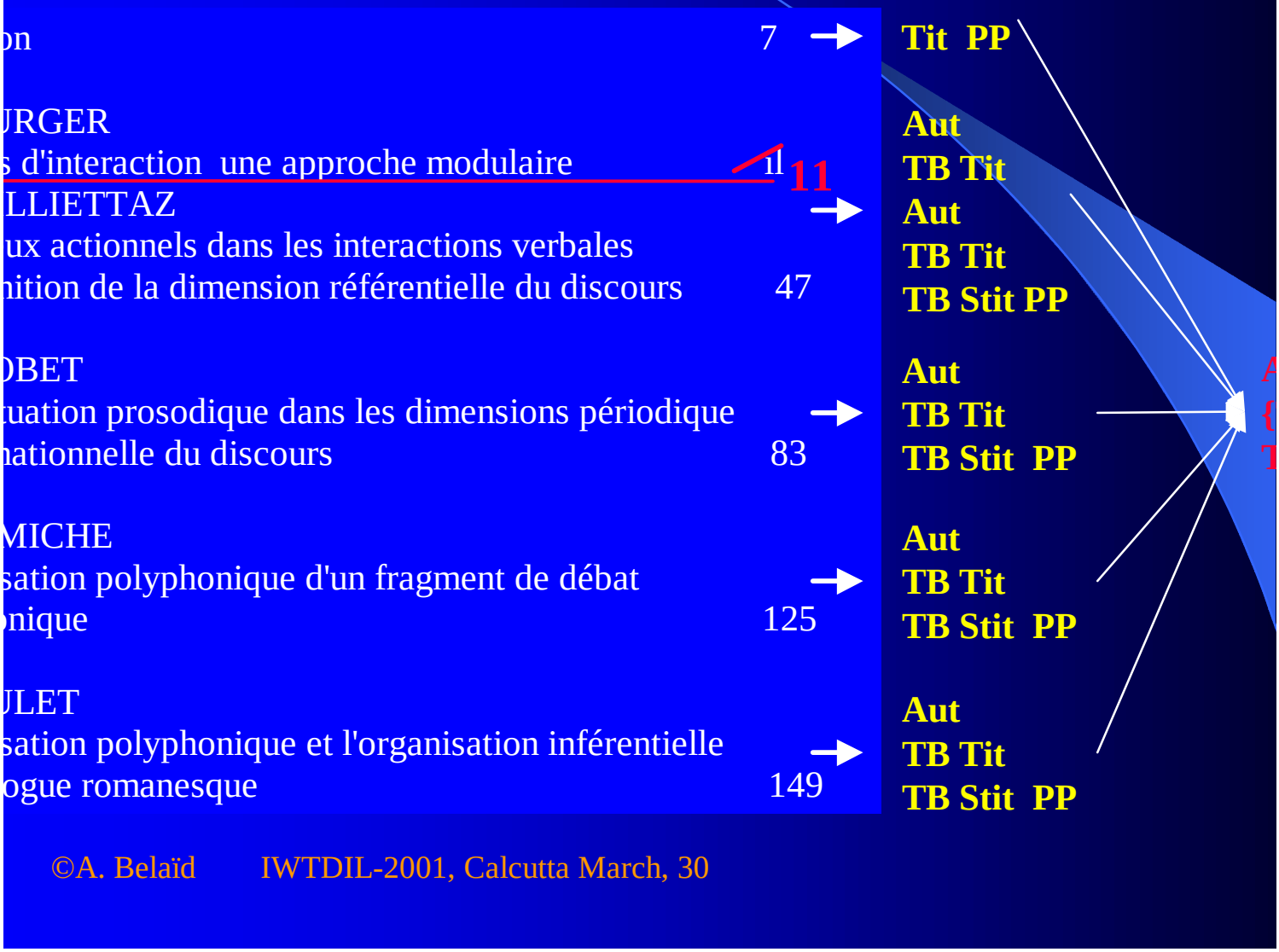
li /NN error /NN measure /NN approach /NN to /TO train /VB feedforwa

neural /NNP networiks /NNP for /IN classification /NN problems /NNS

opapproach /NNP usingspreadsheetsfortheteaching /NNP of /IN statis

probability /VBP

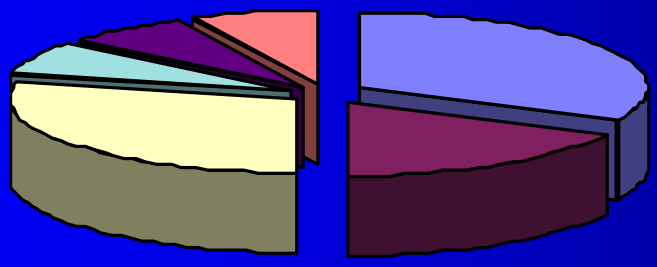
Model Generation



Results (one OCR)

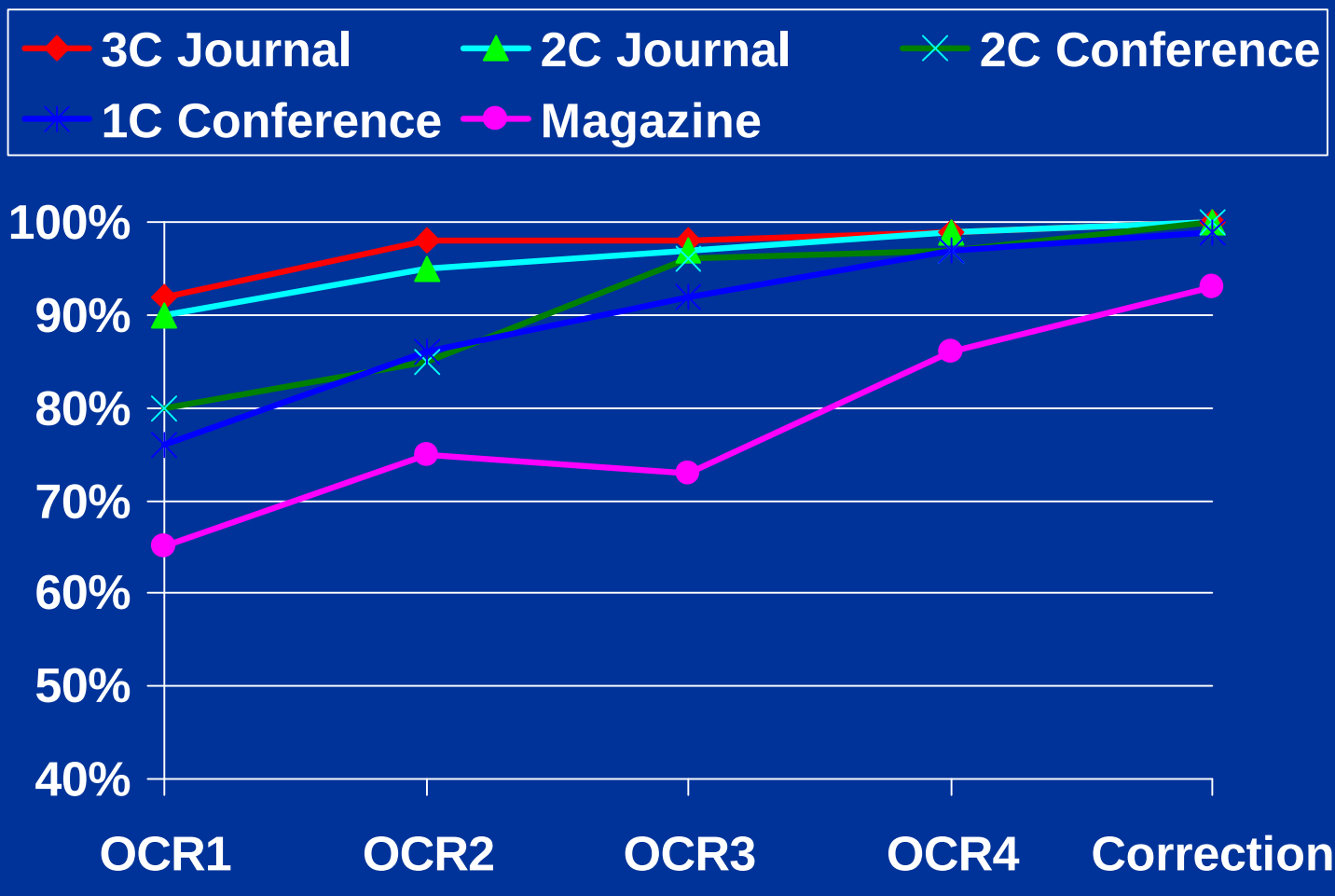
ocument Class	Author Field Number	Identified Authors	Article Number	Identifie Articles
Journal	1037	857	1256	1233
Journal	302	290	326	326
ceedings	204	152	207	194
ceedings	305	227	317	286
Magazine	159	151	168	159
al	2020	1677	2277	2198
		83%		96.5%

Defects

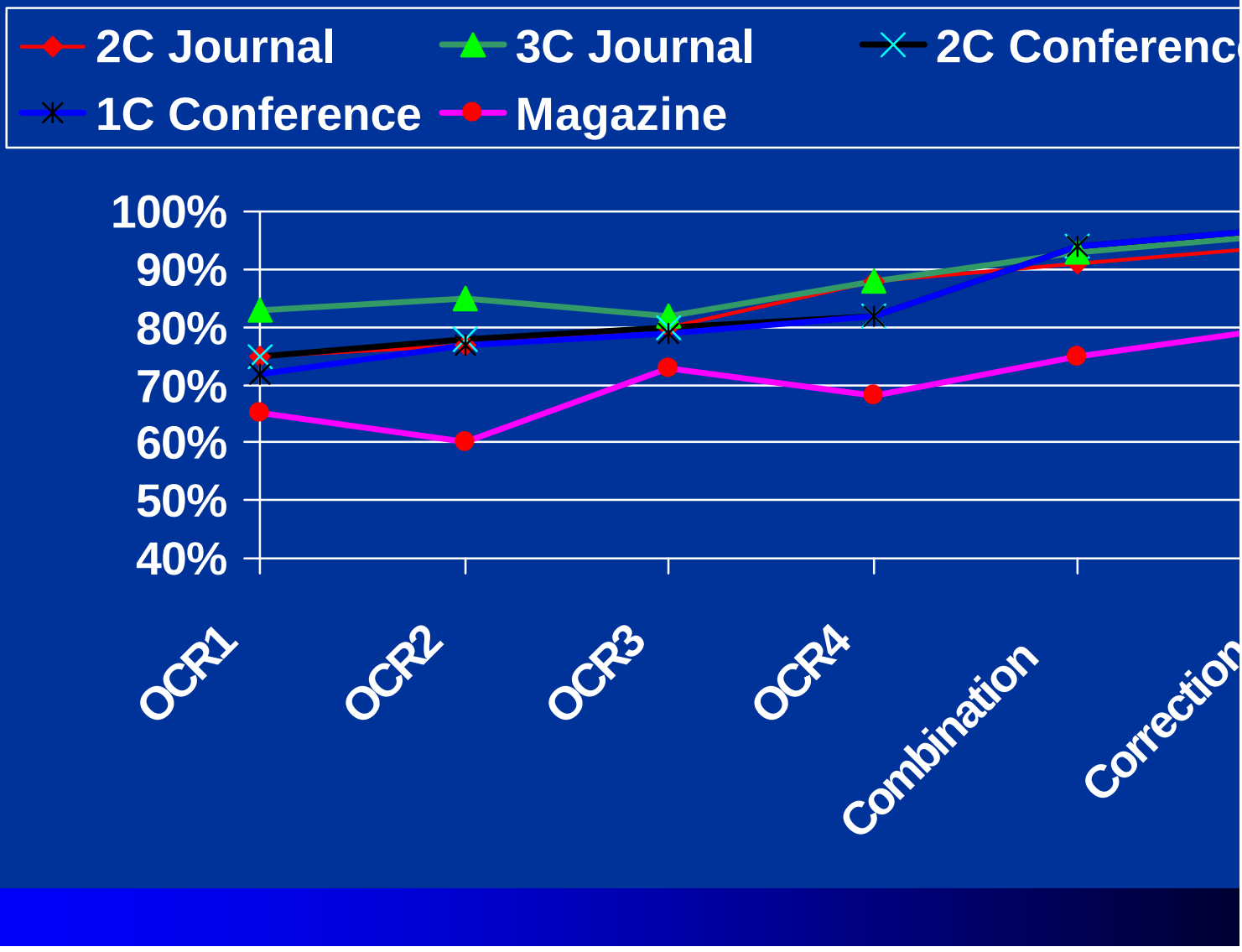


- Page altered
- Page bad or not recognized
- Separator absent or altered
- Non respect of the reference struct
- Error on the last reference
- Section name disturbing

Segmentation



Structure


2C Journal
3C Journal
2C Conferenc
1C Conference
Magazine
100%
90%
80%
70%
60%
50%
40%
OCR1
OCR2
OCR3
OCR4
Combination
Correction

Conclusion and perspective

Prototypes :

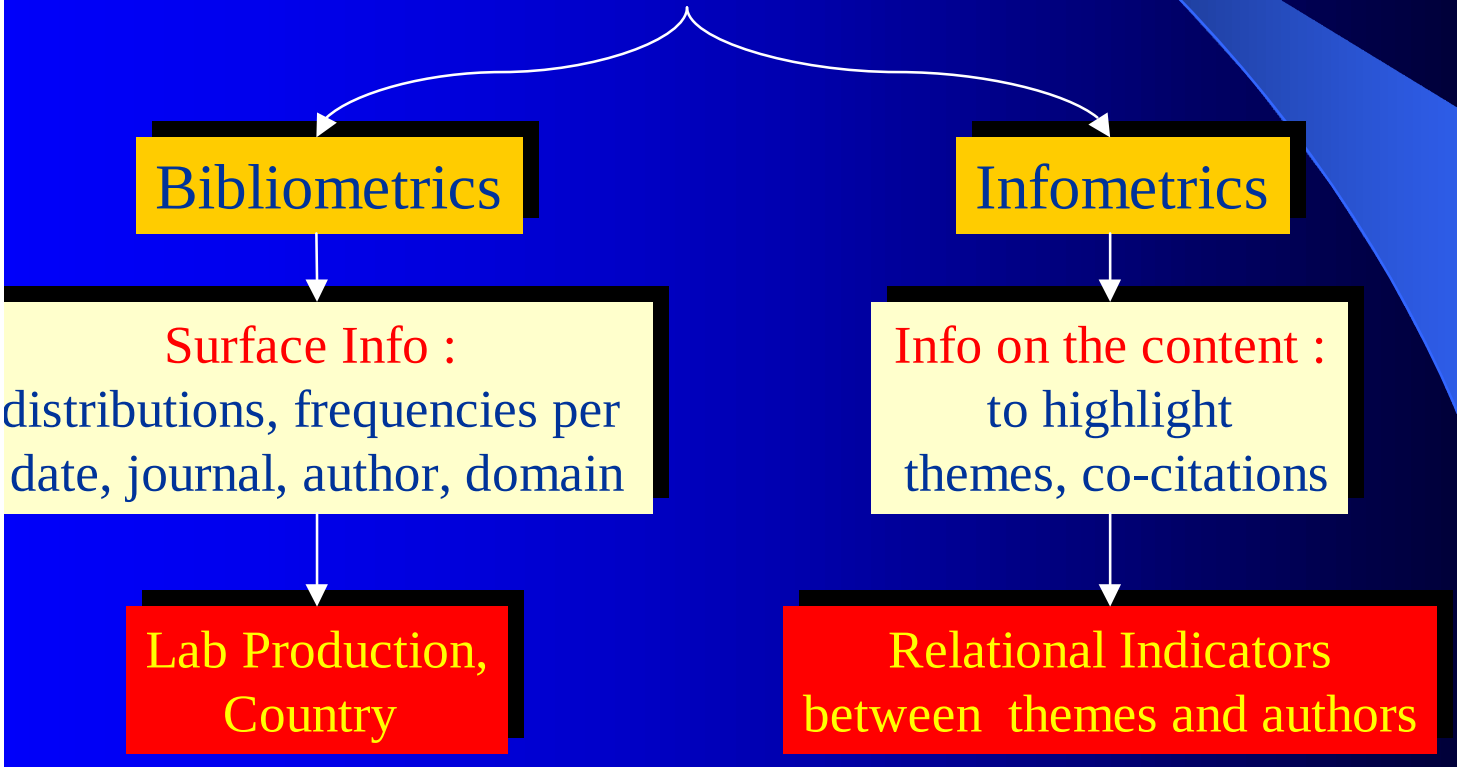
- **INRIA Unit Library connection**
- **Tunisian University Library connection**

Research : Problem Serialization

- Content :
  - OCR Result Improvement
- Linguistics
  - Revise the tag names, more fine
  - Study of other linguistic extension rules

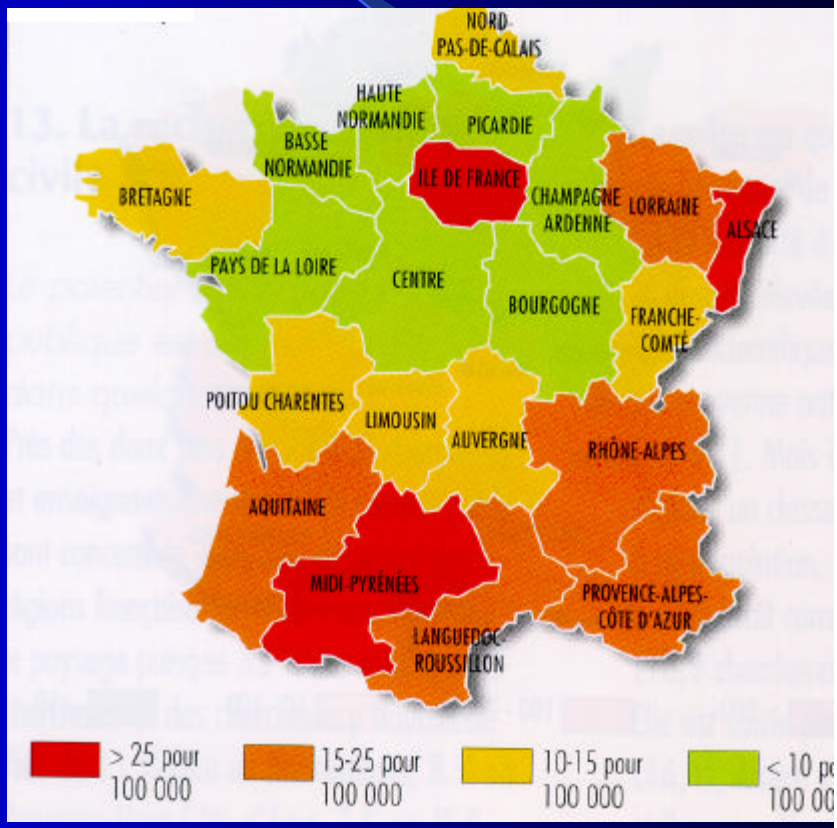
Scientometrics

How to measure the impact of an institution in one domain ?



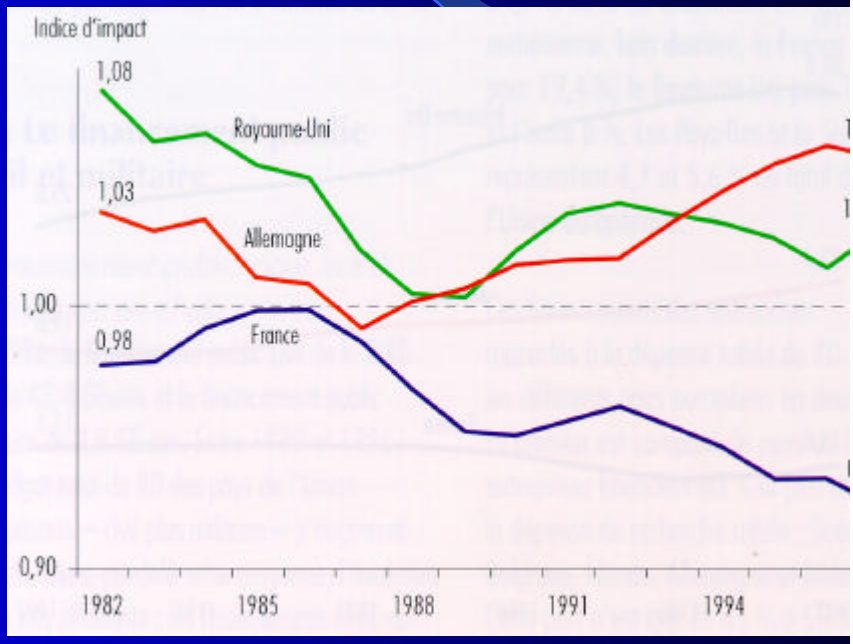
Bibliometrics

- Nb of thesis / population
- Predominance of the Ile de France
- Effects on the neighbor regions



Bibliometrics for article referencing

- mpact Index of the scientific ublications in the E.U.
- mpact = average nb of references er article
-)ne remarks a continuous ecreasing in France, i.e. french rticles are less and less read and eferenced
- everal possible causes...



rying to transform not only
size and position, but also some
using more than one harmonic.
r harmonics can better describe
s of the characters and would
nces to distinguish between
ers having close shapes.

& M Biquard, «Signaux
èmes Linéaires et Bruit en
ronique », Paris, 1992.

M, Lecolinet E, Cheriet M,
. C.Y, Automatic reading of
ve scripts using a reading model
perceptuel concepts, IJDAR,
me1, Number1, 1998, pp 3 - 17.

M.Hamanaka, K.Yamada,
ulumo, « On-line Japanese
acter Recognition Experiments
an Off-line Methid Based on
nalization-cooperated Feauture
action », ICDAR, IEEE, 1993,
04-207.

Authorrs (4) : Côte M., Lecolinet E., Cheriet M., Suen C. Y.
Title : Automatic reading of cursive scripts using a reading model and perceptuel concepts
Editor :

Month : Year : 1998 → Historical
Volume : 1
Number : 1
Publisher :
Address :
Pages : 3-17 → Occurrence + Co-occurrence
Organization :
Booktitle : IJDAR → Indice d'impact
Series :
Note :

The EU Libraries Program

Telematics program : 1990-1994

- **Availability & accessibility** of modern library services
- More rapid penetration of I&CT in a **cost-effective way**
- **Standardization** required for resource sharing
- **Harmonization** and convergence of national policies

28 projects : 28 MECU

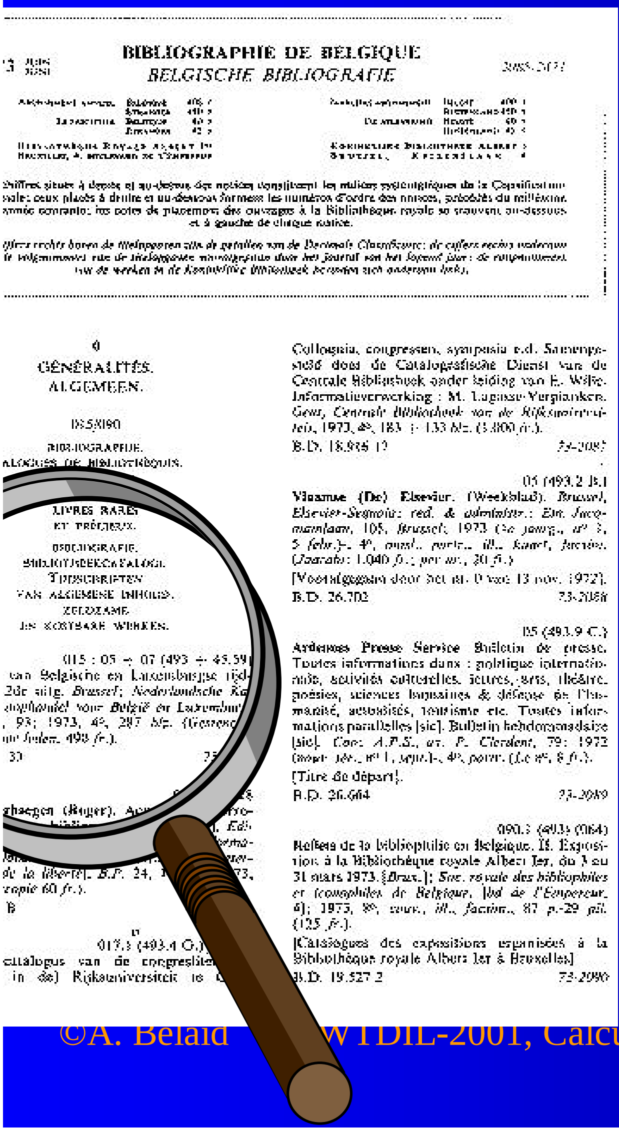
4 Action Lines

- To create, enhance and harmonize machine-readable bibliographies and union catalogues, tools for retrospective conversion of catalogues of important collections
- To impel the international library networking, providin access to source data
- Innovative services for libraries, using new technology
- Development and production of commercially viable product, services and tools specially designed for libra

Three Projects: OCR/ICR

- **BIBLIOTHECA**: Spain + Italy + France
  - Study of a pivot format (XML) for the representation of diffe records
  - Study of a system for indexing and classifying different documents
- **FACIT**: Denmark+Greece+Italy
  - Search for adapted OCR packages for retroconversion with la set of characters, and tools for fast and cheap mass conversion catalogues
- **MORE**: Belgium+France
  - Study of the role & use of dictionaries in the structure modelin and Recognition of catalogues by OCR techniques

Catalogue example



BIBLIOGRAPHIE DE BELGIQUE

BELGISCHE BIBLIOGRAFIE

0

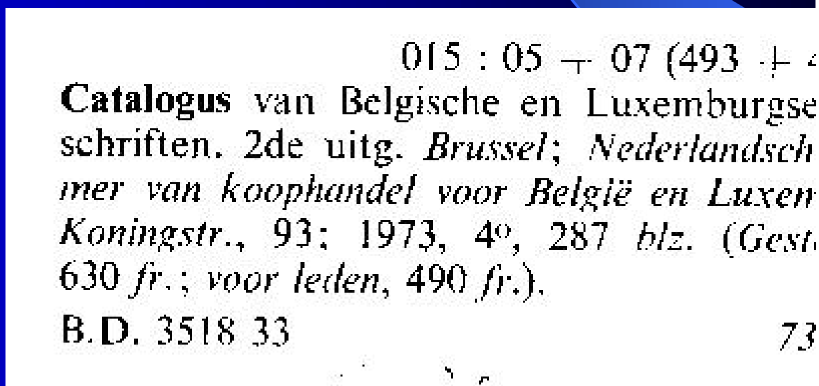
GÉNÉRALITÉS.
ALGEMEEN.

Colloquia, congressen, symposia e.d. Samengesteld door de Catalografische Dienst van de Centrale Bibliotheek onder leiding van E. Wille. Informatieverwerking : M. Laganse-Vergianken. *Gent, Centrale Bibliotheek van de Rijksuniversiteit*, 1973, 4°, 183 + 133 *blz.* (3.800 *fr.*).
B.D. 18.935 12 — *73-2087*

05 (493.2 B.)
Vlaamse (De) Elsevier. (Weekblad). *Brussel, Elsevier-Sequoia: red. & administr.: Em. Jacqmainlaan, 105, Brussel*; 1973 (4e jaarg., nr 3, 5 febr.)-. 4°, omsl., portr., ill., kaart, facsim. (*Jaarabo.*: 1.040 *fr.*; per nr., 30 *fr.*)
[Voortgezet door het nr. 0 van 13 nov. 1972].
B.D. 26.702 — *73-2088*

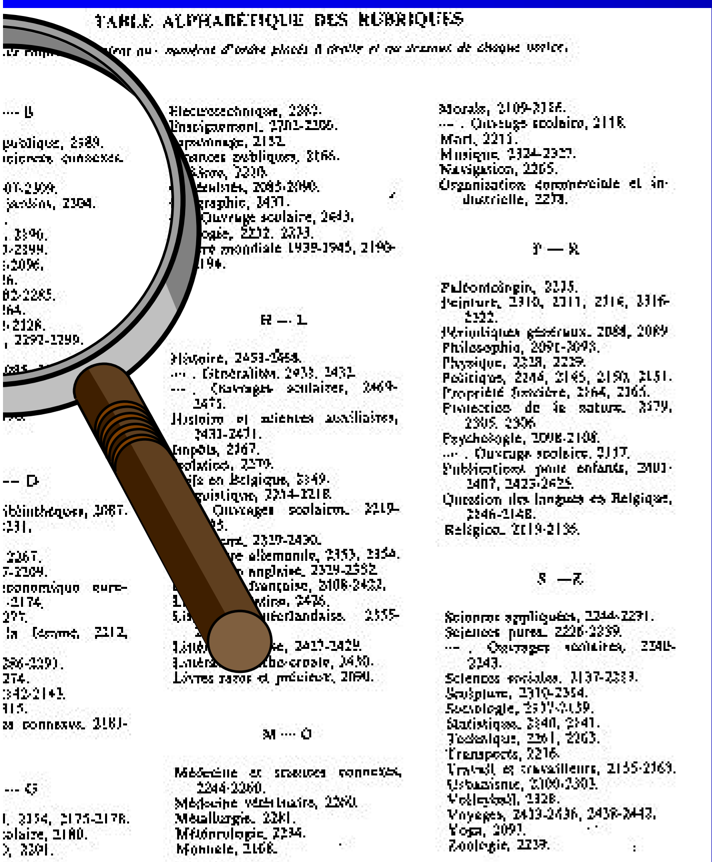
05 (493.9 C.)
Ardennes Presse Service. Bulletin de presse. Toutes informations dans : politique internationale, activités culturelles, lettres, arts, théâtre, poésies, sciences humaines & défense de l'humanité, actualités, tourisme etc. Toutes informations parallèles [sic]. Bulletin hebdomadaire [sic]. *Caen: A.P.S., av. P. Clerdent, 79*: 1972 (nouv. sér., nº 1, sept.)-. 4°, portr. (*Le nº, 8 fr.*).
[Titre de départ].
B.D. 26.664 — *73-2089*

090.3 (493) (064)
Reflets de la bibliophilie en Belgique. II. Exposition à la Bibliothèque royale Albert Ier, du 3 au 31 mars 1973. [*Brux.*]; *Soc. royale des bibliophiles et iconophiles de Belgique*, [*bd de l'Empereur*, 4]; 1973, 8°, couv., ill., facsim., 87 p.-29 pl. (125 *fr.*).
[Catalogues des expositions organisées à la Bibliothèque royale Albert Ier à Bruxelles]
B.D. 19.527 2 — *73-2090*



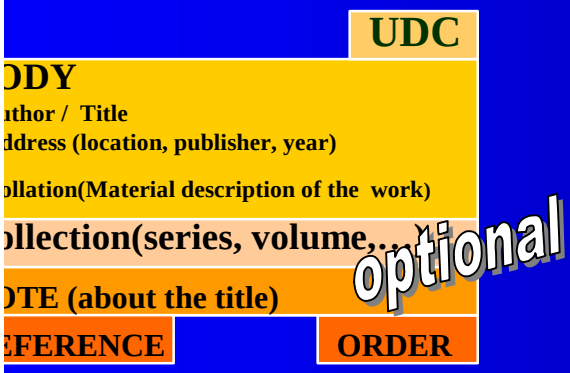
015 : 05 + 07 (493 + 4
Catalogus van Belgische en Luxemburgse
schriften. 2de uitg. *Brussel*; *Nederlandsch*
mer van koophandel voor België en Luxem
Koningstr., 93; 1973, 4°, 287 *blz.* (*Gest*
630 *fr.*; *voor leden*, 490 *fr.*).
B.D. 3518 33 — *73*

Catalogue Indexes

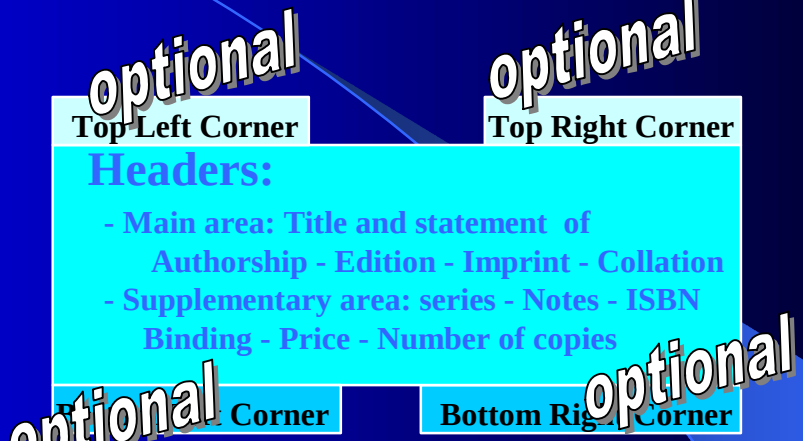


Administration publique, 2189
Agriculture et sciences complexes 2266, 2267
Architecture, 2307-2309
Architecture de jarsdins, 2304
Arpentage, 2227
ASSistance, 2195, 2196 ...

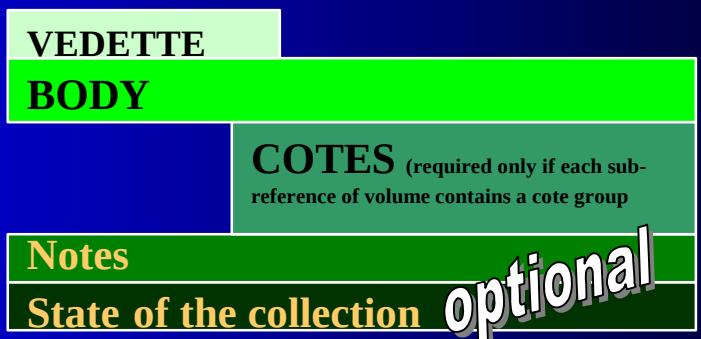
Record structures



Belgian Library



Danish Library



French Library

Application on Belgian records

159.962

ir (Gérard). Je suis fakir. ([Par] Gé-
-Belair). (*Verviers, Editions Gérard &*
32° carré, couv., ill., 158 *p.* (30 *fr.*).
-flash, 352).
ductif : Souvenirs, révélations, con-

4 352 73-2108

Pre-ISBD
~ 1973

UDC	<675 I=bb <$a 159.962</$a></675>
Heading	<200 I=0b <$f Gérard Liger-Belair</
	<$a Je suis fakir </$a></20
Author	<700 I=b0 <$a Liger-Belair</$a>
	<$b Gérard <$b></700>
Publisher	<210 I=bb <$a Verviers</$a>
	<$c Editions Gérard & C°
	<$d [1973]</$d></210>
Collation	<215 I=bb <$d 320 carré</$d>
	<$c couv., ill.</$c>
	<$a 158 p.</$a></215>
Price	<010 I=bb <$d 30 BEF</$d></010>
Collection	<225 I=2b <$a Marabout-flash
	<$v 352</$v></225>
Note	<517 I=0i1 <$a Souvenirs, révélations
	conseils</$a></517>
Ref.	<900 I=bb <$a B.D. 14.814352</$a>
	<$b 73-2108</$b></900>

Extension to bibliographic references

] J. Juan, Y. Y. Tang, and C. Y. Suen. Four Directional Adjacency Graphs (fdag) a
tion in Locating Fields in Forms. In Third International Conference on Document
ognition (ICDAR'95), pages 752-755. IEEE Computer Society Press, Aug. 1995.

: J. Juan, Y. Y. Tang, and C. Y. Suen
: Four Directional Adjacency Graphs (fdag) and their Application in Locating fi
:
: Aug
: 1995
:
:
: IEEE Computer Society Press
:
: 752-755
n:
: Third International Conference on Document Analysis and Recognition (ICDA
:
: