

Vocabularies Support for tracing Knowledge Diversity

Dr.Devika P. Madalli

Indian Statistical Institute
Bangalore, INDIA



Indian Statistical Institute

- Established in: 1931
- Institute of National Importance: 1959
- First to commission a computer in India
- Founder: Prof. P.C. Mahalanobis



DRTC



Documentation Research and Training Centre

- Established: 1962
- Founder: Prof. S.R. Ranganathan

S.R. Ranganathan

- Father of Indian Library Science
- Father of Faceted Classification (ontology)
- Creator of Colon Classification
- Creator of Classified Catalogue Code
- Creator of Chain Indexing (followed by BNB)

Research Areas

- Natural Language Processing
- Quantitative Methods in LIS
- Information Retrieval and Data Mining
- Knowledge Management
- Digital Libraries
- Multi-Lingual Information Systems
- Classification (Ontologies)

Software Developed

Manu	For thesaurus construction
Prometheus	POPSI based index generation
Panizzi	For automatic identification of Bibliographic data elements from the title page
Viswamitra & Vyasa	Automatic construction of Call Numbers, maintenance of Schedules, indexes, etc
Pygmalion	Packages for retro-conversions
Ekalavya	Computer aided teaching packages

DL Test-beds



Eprints 2



Fedora



CDSWare



Green Stone Digital Library

LDL : Librarians' Digital Library

<https://drtc.isibang.ac.in>

powered by



Communities & Collections

Library and Information Science

- [Publications / Articles](#)
- [Theses / Dissertations](#)
- [PowerPoint Presentations](#)
- [Demo of Multilingual Documents](#)
- [Photographs of LIS activities](#)
- [Photographs of S.R. Ranganathan](#)

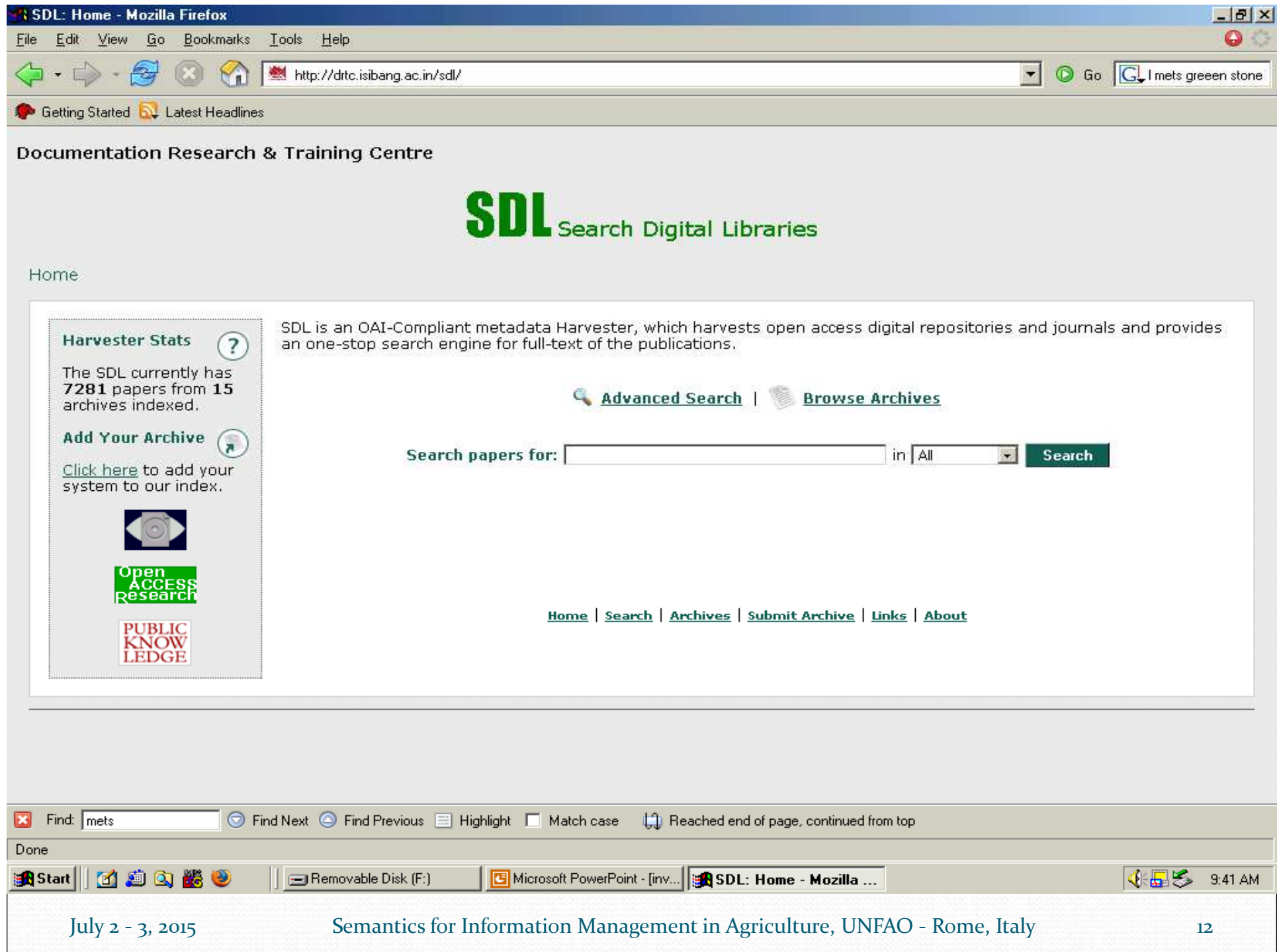
Membership From...

- India
- USA
- France
- UK
- South Africa
- Thailand
- Austria
- Italy



Harvester Service

<http://drtc.isibang.ac.in/sdl>



SDL: Browse Archives - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://dlc.isibang.ac.in/sdl/archives.php

Go I mets green stone

Getting Started
 Latest Headlines

Home > Browse Archives

Select an archive to browse...

▶ [All Archives](#) (7281 records)

▶ [Australian Library and Information Science Association \(ALIA\)](#) (19 records)

▶ [CaltechLIB: Caltech Library System Papers and Publications, USA](#) (34 records)

▶ [CCSD: Sciences de l'Information et de la Communication, France](#) (606 records)

▶ [CNR Bologna Research Library, Italy](#) (17 records)

▶ [D-Lib Magazine](#) (319 records)

▶ [Diálogo Científico utiliza](#) (201 records)

▶ [DLIST, University of Arizona, USA](#) (486 records)

▶ [E-LIS: E-Prints in Library and Information Science](#) (2825 records)

▶ [INFLIBNET, India](#) (428 records)

▶ [Instituto Brasileiro de Informação](#) (431 records)

▶ [LDL: Librarians' Digital Library](#) (188 records)

▶ [OCLC](#) (349 records)

▶ [OpenMed, NIC, India](#) (509 records)

▶ [University of North Carolina, USA](#) (365 records)

▶ [WWW Conference Archive EPrint servers](#) (504 records)

Quick Search

in All

Search

[Advanced Search](#)

Find: Find Next Find Previous Highlight Match case Reached end of page, continued from top

Done

Start

Removable Disk (F:)
 Microsoft PowerPoint - [inv...]
 SDL: Browse Archive...

9:41 AM

July 2 - 3, 2015

Semantics for Information Management in Agriculture, UNFAO - Rome, Italy

13

Discussion Forum

DLRG: Digital Library Research Group

- Presently over 250 members
- <http://drtc.isibang.ac.in/dlrg>



Indus

(a DSpace based harvester)

- 48 Asian Countries
- 26 Countries have repositories (openDOAR)
- Around one third of them have exclusive Agricultural repositories
- More OAI-based Agri. Journals

<http://drtc.isibang.ac.in/indus>

Indus

- Indus covers both repositories and OAI based Journals
- Presently
 - *About 10 countries repositories are harvested*
 - *57 Journals on Agriculture*
 - *8 Digital Repositories*
 - *About 50k records*

Indus Home - Mozilla Firefox

Indus Home

drtc.isibang.ac.in/indus/

Google

Indus

Promoting data sharing and development of trust in agricultural sciences

Indus Home

DSpace Repository

DSpace is a digital service that collects, preserves, and distributes digital material. Repositories are important tools for preserving an organization's legacy; they facilitate digital preservation and scholarly communication.

Repositories and Journals in Indus

Select a Repository/Journal to browse

- [Bangladesh](#)
- [India](#)
- [Indonesia](#)
- [Iran](#)
- [Malaysia](#)
- [Philippines](#)
- [Sri Lanka](#)
- [Thailand](#)
- [United Arab Emirates](#)

Search Indus

Enter some text in the box below to search Indus.

Go

Search Indus

Go

[Advanced Search](#)

Browse

All of Indus
[Repositories & Journals](#)
[By Issue Date](#)
[Authors](#)
[Titles](#)
[Subjects](#)

My Account

[Login](#)
[Register](#)

Discover

Author
[IPS, Editor \(343\)](#)
[International Rice Research Institute. \(76\)](#)
[International Rice Research Institute \(65\)](#)
[Anon \(52\)](#)
[KRISHNAMOORTHY, B \(52\)](#)
[ANANDARAI, M \(51\)](#)



Work on Vocabularies

July 2 - 3, 2015

Semantics for Information Management in
Agriculture, UNFAO - Rome, Italy

18

GeowordNet

-- Biswanath Dutta, Fausto Giunchiglia, Vincenzo Maltese

- Space and Time are the two fundamental dimensions of the universe of knowledge
- Space is essential to understand the physical universe
 - by “Space”, it is meant, *surface of the earth, the space inside it and the space outside it*
 - it can be interpreted by its geographical features including others like, buildings and other man-made structures

Issues

- There is a need for supporting semantic interoperability between people and also between applications
- Definition of entity types and corresponding properties have become a central issue in data exchange standards
- Current standards do not address the actual semantic interoperability problem
 - mainly aim at *syntactic* agreement by fixing the standard terms

Approach

- GeoWordNet*, a multi-lingual ontology that overcomes the qualitative and quantitative limitations over previous ontologies
- It is based well founded methodologies and guiding principles for developing the faceted ontologies

*a subset of GeoWordNetis available as open source in plain CSV and RDF formats and can be downloaded from:

<http://geowordnet.semanticmatching.org/>

Main contribution

- We proposed here a **methodology** and a **limited set of guiding principles** to construct geo-spatial ontology
- They are based on the notion of **facet** and **analytico-synthetic approach** borrowed from Library Science

Facet

[First Introduced by Ranganathan (1930s) in Library and Information Science]

- “A generic term used to denote any component – be it a basic subject or an isolate – of a compound subject, ...” - Ranganathan
- It is a category that expresses some **aspect** of the knowledge being described
- A facet is a hierarchy of homogeneous terms, where each term in the hierarchy denotes a primitive atomic **concept**
- **E.g.**, Organ facet, geographical facet, language facet, property facet, author facet, religion facet, commodity facet, etc.

Facet Example:

Language

by Indo-European

Teutonic

Gothic

English

American English

German

Latin

Italian

French

Greek

by Dravidian

Tamil

Tulu

by Geographic location

Asian language

(collective treatment)

Japanese language

Indian language

African language

Step 1: identification of the atomic concepts

- Some of the relevant sub-trees in WordNet are:
 - location
 - artifact, artefact
 - body of water, water
 - geological formation, formation
 - land, ground, soil
 - land, dry land, earth, ground, solid ground, terra firma

Note: not necessarily all the nodes in these sub-trees need to be part of the space domain. For example, the **descendants of artifact**, like, *article*, *anachronism*, *block*, etc. are not.

Analysis

Mountain

- the well defined **elevated land**
- formed by the **geological formation** (where geological formation is a natural phenomenon)
- **altitude** in general >500m

Hill

- the well defined **elevated land**
- formed by the **geological formation**, where geological formation is a natural phenomenon
- **altitude** in general <500m

Stream

- a **body of water**
- a **flowing** body of water
- **no** fixed boundary
- confined within a bed and stream banks

River

- a **body of water**
- a **flowing** body of water
- **no** fixed boundary
- confined within a bed and stream banks
- **larger** than a brook

26

Synthesis

Body of water

Flowing body of water

Stream

Brook

River

Stagnant body of water

Pond

Landform

Natural depression

Oceanic depression

Oceanic valley

Oceanic trough

Continental depression

Trough

Valley

Natural elevation

Oceanic elevation

Seamount

Submarine hill

Continental elevation

Hill

Mountain

* each term in the above has gloss and is linked to synonym(ous) terms in the knowledge base

Facets and sub-facets

- Space [**Domain**]
 - by geographical features [**Entity types**]
 - by water formation
 - by land formation
 - by land
 - by administrative division
 - ...
 - by relations [**Relation**]
 - spatial relation
 - direction, internal, external, longitudinal, sideways, etc.
 - functional relation (e.g., primary inflow, primary outflow)
 - ...
 - by property [**Attribute**]
 - latitude
 - Longitude
 - dimension
 - ...

28

Vocabularies to trace Knowledge Diversity

➤ Living Knowledge Project

➤ **FP7 FET project**

<http://livingknowledge.europarchive.org/>

Challenges

- IR Challenges in general
 - High recall
 - Low precision
 - Natural language processing
 - Disambiguation problems
 - E.g., a word “bass”
 - Sense 1: A kind of saltwater fish
 - Sense 2: Tones of low frequency
- Natural language sentences:
- I went fishing for some sea bass
 - The bass line of the song is too weak

Solution

- A Large Scale, Domain Specific LR based on Facet based KO is a better Resource for addressing the challenges of Low Precision and High Recall

Resources

- Language resources
 - General purpose language resources
 - WordNet (<http://wordnet.princeton.edu/>)
 - MultiWordNet (<http://multiwordnet.fbk.eu/english/home.php>)
 - EuroWordNet (<http://www.illc.uva.nl/EuroWordNet/>)
 - Rogets's thesaurus
 - Domain specific language resources
 - Dewey Decimal Classification (DDC)
 - Library of Congress Classification (LCC)
 - Universal Decimal Classification (UDC)
 - Bliss Bibliographic Classification (BC)
 - Colon Classification (CC)
 - AGROVOC
 - Art and Agriculture Thesaurus

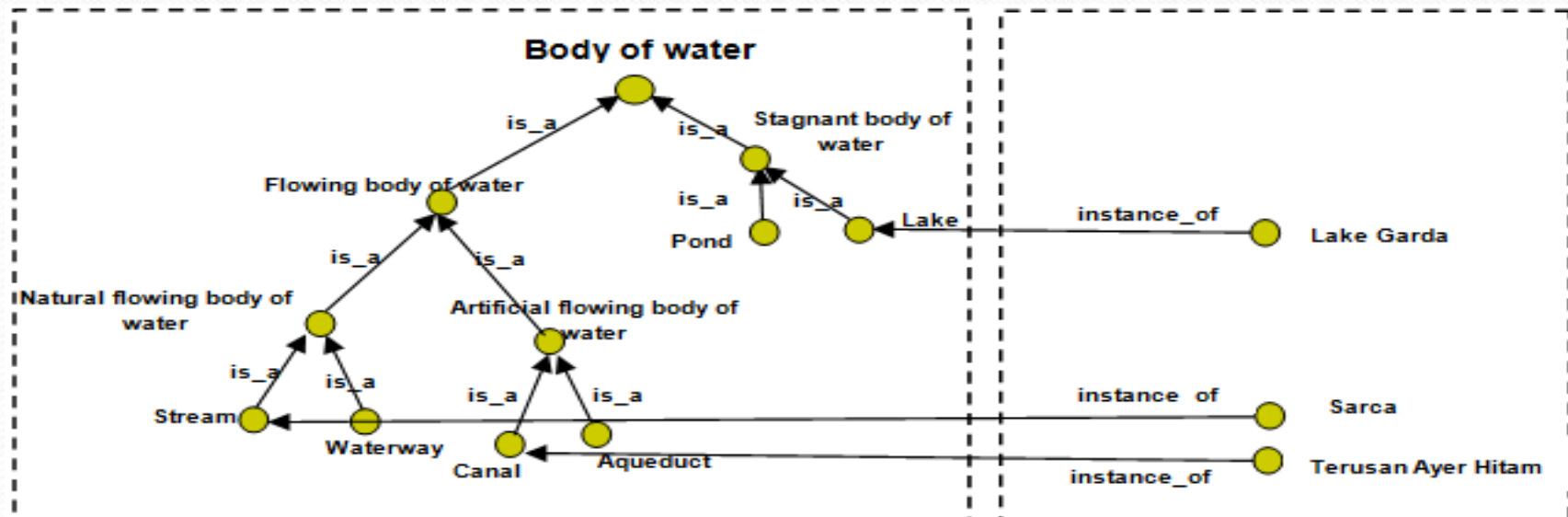
DERA

[F. Giunchiglia and B. Dutta, 2011]

- Consists of:
 - Domain [D]
 - Entity [E]
 - Relation [R]
 - Attribute [A]
- It is a further refined and simplified form of Bhattacharyya's DEPA
- Has direct mapping to DL
- Emphasis is on the **named entities**

Entity

- ❑ *An elementary component that consists of classes (categories) and their instances, having either perceptual correlates or only conceptual existence in a domain in context*
- ❑ $E = \langle \{e\}, \{E\} \rangle$
- ❑ e = Entity class - consists of the core classes within a domain
- ❑ E = Entity - consists of the real world (named) entities which are instances of the entity classes “ e ”



Attractiveness of Photos

- Community-based models for classifying/ranking images according to their appeal. [WWW09]

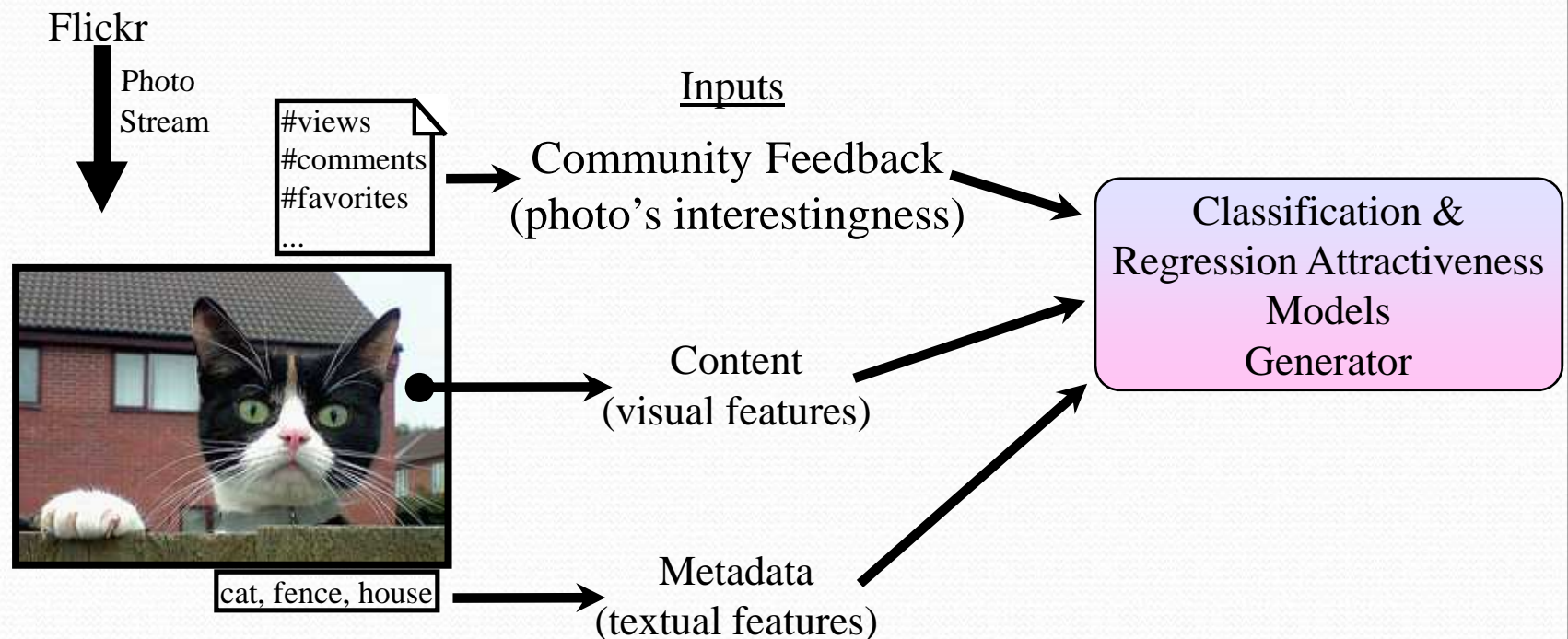


Photo Annotation

Modelling image content as bags-of-visual-terms
learnt through hierarchical K-means clustering

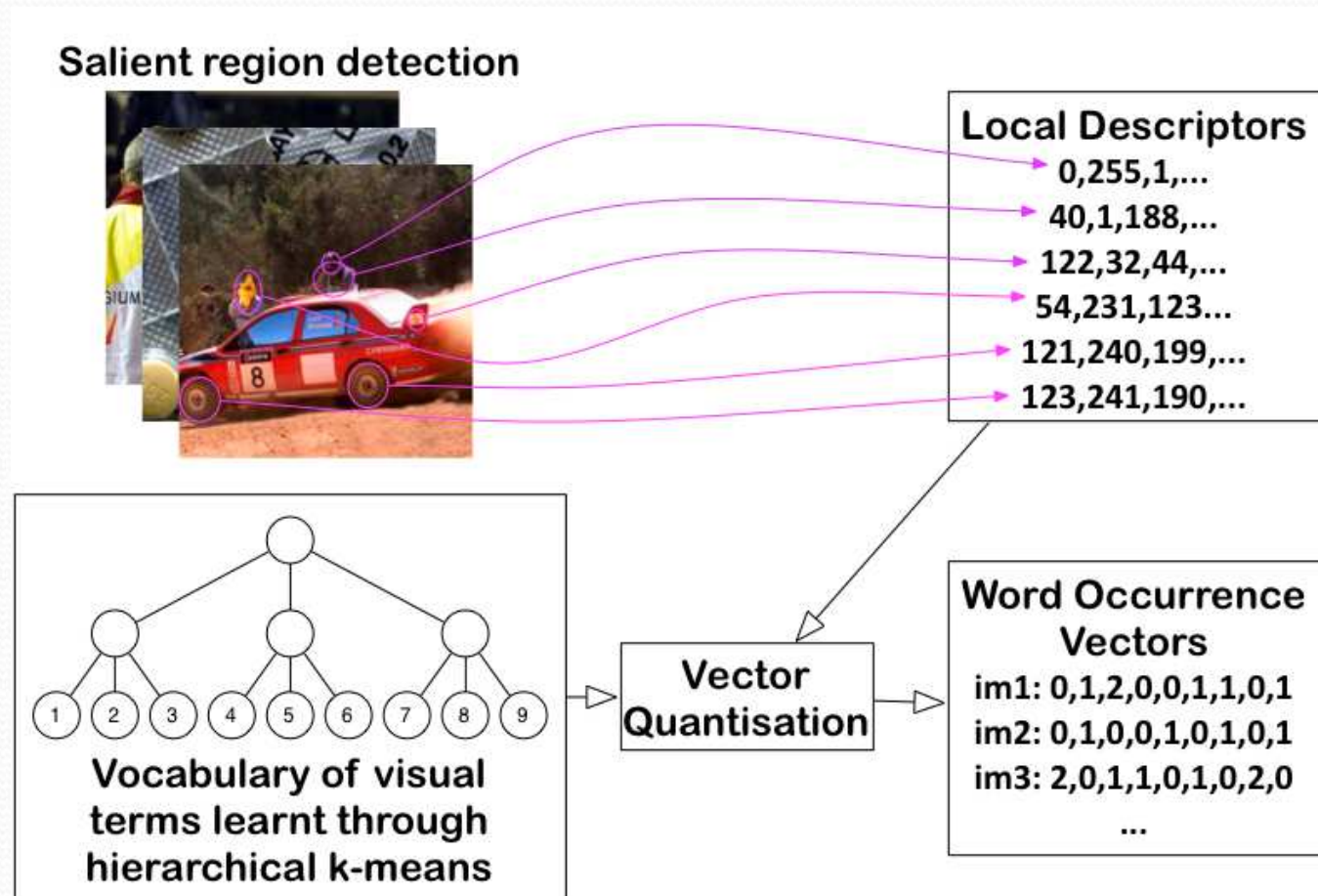
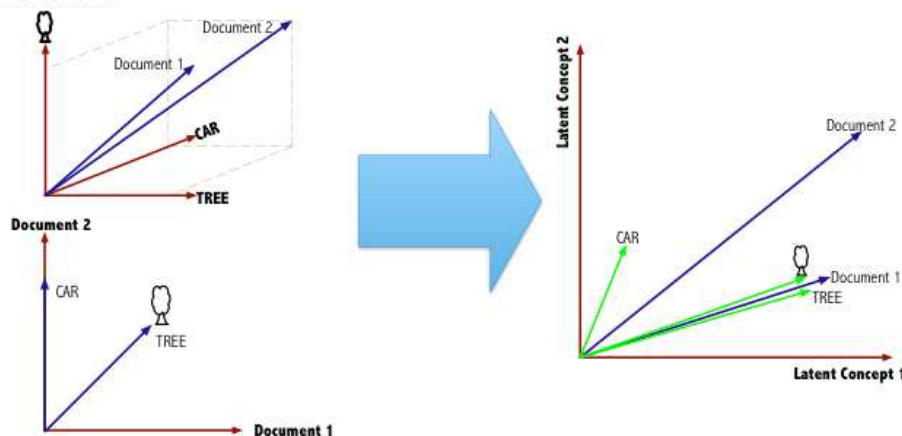
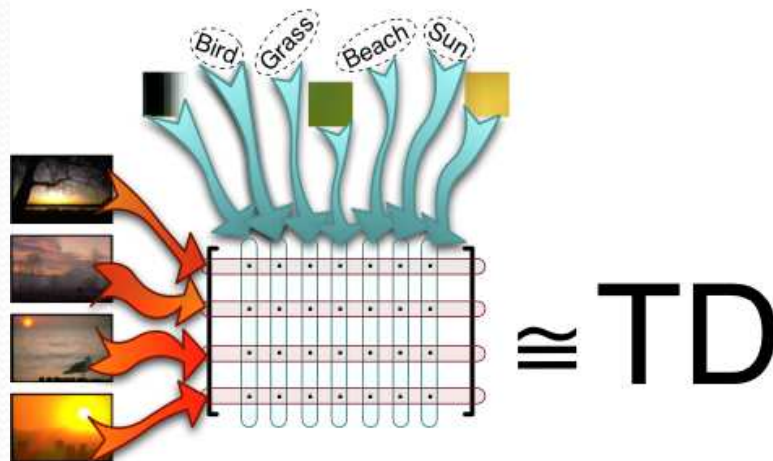


Photo Annotation

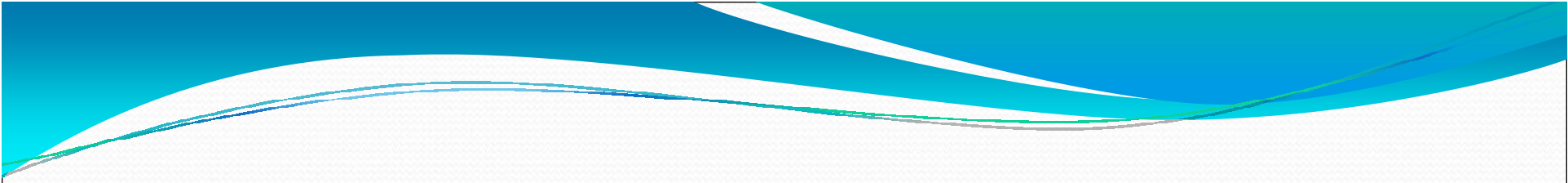
Automatically annotating and classifying images using a *semantic space* approach.



Run	Average EER	Average AUC
ISIS University of Amsterdam_76_2_1245075326057.txt	0.234476	0.838699
ISIS University of Amsterdam_76_2_1245075554764.txt	0.234476	0.838699
ISIS University of Amsterdam_76_2_1245075919629.txt	0.235479	0.837531
ISIS University of Amsterdam_76_2_1245076767051.txt	0.243547	0.830005
LEAR_44_2_1245582451309.txt	0.249469	0.823105
ISIS University of Amsterdam_76_2_1245076320558.txt	0.252997	0.821731
CVIUI2R_22_2_1244628714641.txt	0.253296	0.813893
FIR2_92_2_1245516363637.txt	0.253566	0.817159
FIR2_92_2_1245417583652.txt	0.253572	0.817153
FIR2_92_2_1245516525876.txt	0.253685	0.816811
CVIUI2R_22_2_1244629050173.txt	0.255945	0.811421
LEAR_44_2_1245582143586.txt	0.256169	0.804713
LEAR_44_2_1245581505805.txt	0.258642	0.8133
XRCE_36_2_1245429345115.txt	0.267301	0.802704
FIR2_92_2_1245516822169.txt	0.271969	0.762276
LEAR_44_2_1245581967963.txt	0.273385	0.79392
bpacad_18_2_1245622473940.txt	0.291718	0.773133
bpacad_18_2_1245622367725.txt	0.296315	0.771124
bpacad_18_2_1245623805085.txt	0.304113	0.7463
LEAR_44_2_1245581860906.txt	0.304383	0.756649
MMIS_33_2_1245434554581.txt	0.312366	0.744231
bpacad_18_2_1245622065724.txt	0.322632	0.733264
IAM Southampton_30_2_1245438072355.txt	0.3304	0.71483
ISIS_14_2_1245595947674.txt	0.330819	0.720931
IAM Southampton_30_2_1245519187248.txt	0.33345	0.71157
IAM Southampton_30_2_1245519327555.txt	0.34212	0.69977
bpacad_18_2_1245621915717.txt	0.346106	0.70786
MMIS_33_2_1245586552541.txt	0.352478	0.68941
MMIS_33_2_1245611281967.txt	0.352485	0.689407
MMIS_33_2_1245674693001.txt	0.352612	0.689342
MMIS_33_2_1245601239738.txt	0.356945	0.684821
ISIS_14_2_1245664294478.txt	0.360759	0.685542
LIP6_17_2_1245498965150.txt	0.372169	0.673089
MRIM_50_2_1245578233616.txt	0.38363	0.64345
LIP6_17_2_1245579903668.txt	0.383789	0.651276
MRIM_50_2_1245574356224.txt	0.392492	0.592542
LIP6_17_2_1245610812916.txt	0.406509	0.629883
LIP6_17_2_1245618383113.txt	0.410032	0.622688
ISIS_14_2_1245664386697.txt	0.428975	0.600487
MRIM_50_2_1245574874836.txt	0.439504	0.573579
AVEIR_81_2_1245664654451.txt	0.440589	0.550866
MRIM_50_2_1245573948330.txt	0.442568	0.528129
Wroclaw University of Technology_49_2_1245624864334.txt	0.446024	0.220957
Wroclaw University of Technology_49_2_1245622900052.txt	0.447512	0.214875
Wroclaw University of Technology_49_2_1245622952589.txt	0.447512	0.214875
Wroclaw University of Technology_49_2_1245622279557.txt	0.451702	0.18037
KameyamaLab_21_2_1245594455534.txt	0.452374	0.164048
AVEIR_81_2_1245664820829.txt	0.454191	0.566118
Wroclaw University of Technology_49_2_1245621871943.txt	0.454221	0.17054
KameyamaLab_21_2_1245617918119.txt	0.455052	0.15604
KameyamaLab_21_2_1245617609041.txt	0.457482	0.152011
AVEIR_81_2_1245664820829.txt	0.457482	0.152011
AVEIR_81_2_1245664820829.txt	0.457482	0.152011
Ka	0.457482	0.152011
UA	0.457482	0.152011
LS	0.457482	0.152011
ap	0.457482	0.152011
IN	0.457482	0.152011
AV	0.457482	0.152011
IN	0.457482	0.152011
IN	0.457482	0.152011
LS	0.457482	0.152011
ap	0.457482	0.152011
ap	0.457482	0.152011
IN	0.457482	0.152011
IN	0.457482	0.152011
Ka	0.457482	0.152011
LIP	0.457482	0.152011
Ra	0.457482	0.152011
CE	0.457482	0.152011
CE	0.457482	0.152011
CE	0.457482	0.152011
TE	0.457482	0.152011
TELECOM ParisTech_39_2_1245415171956.txt	0.527192	0.458622

Overall Result:

- * Competitive performance
- * Low computational complexity compared to other entries

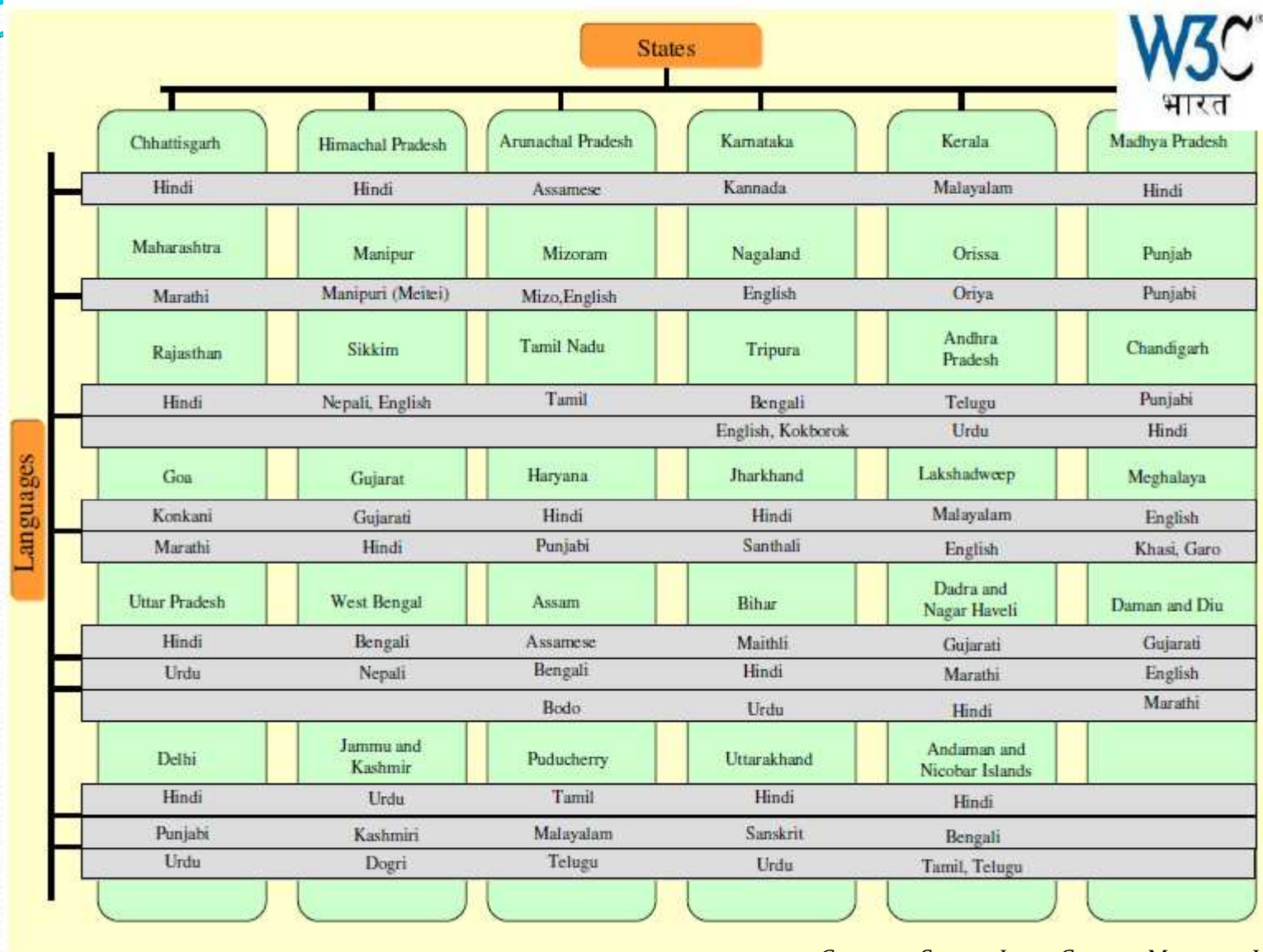


Multilingual Complexity In India

Languages of India

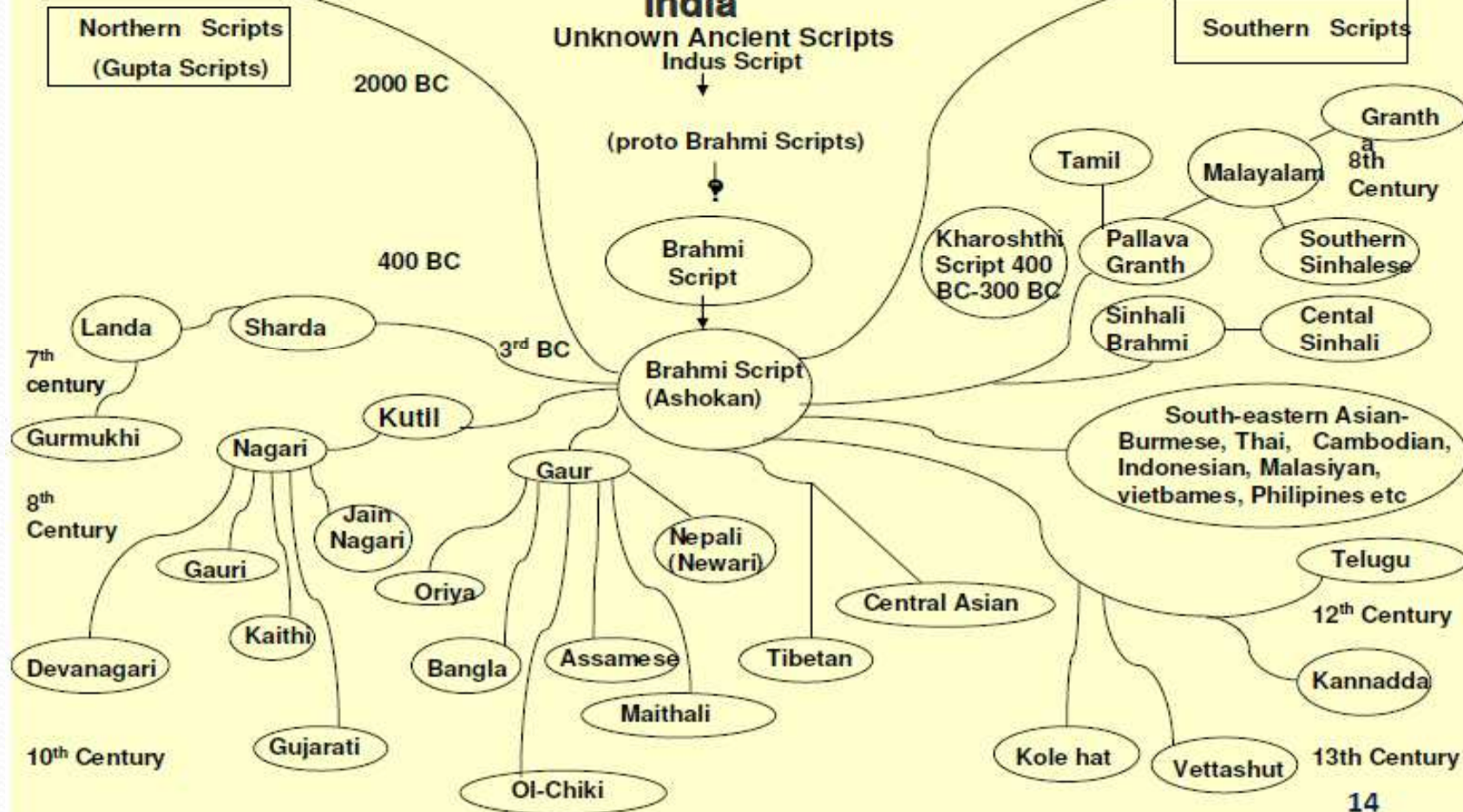
- According to Census 2001 India has 122 major languages and 2371 dialects.
- Out of 122 languages 22 are constitutionally recognized languages.
- Linguistic Diversity is very rich and wide in India
- One Language –many scripts
- Many Language –one script
- Culturally different depending on region though using same script for different languages.
- Even wide difference for same language across different parts of the country

~ Courtesy: Swaran Lata (DIT) , Country Manager , W3C India



~ Courtesy: Swaran Lata , Country Manager , W3C India

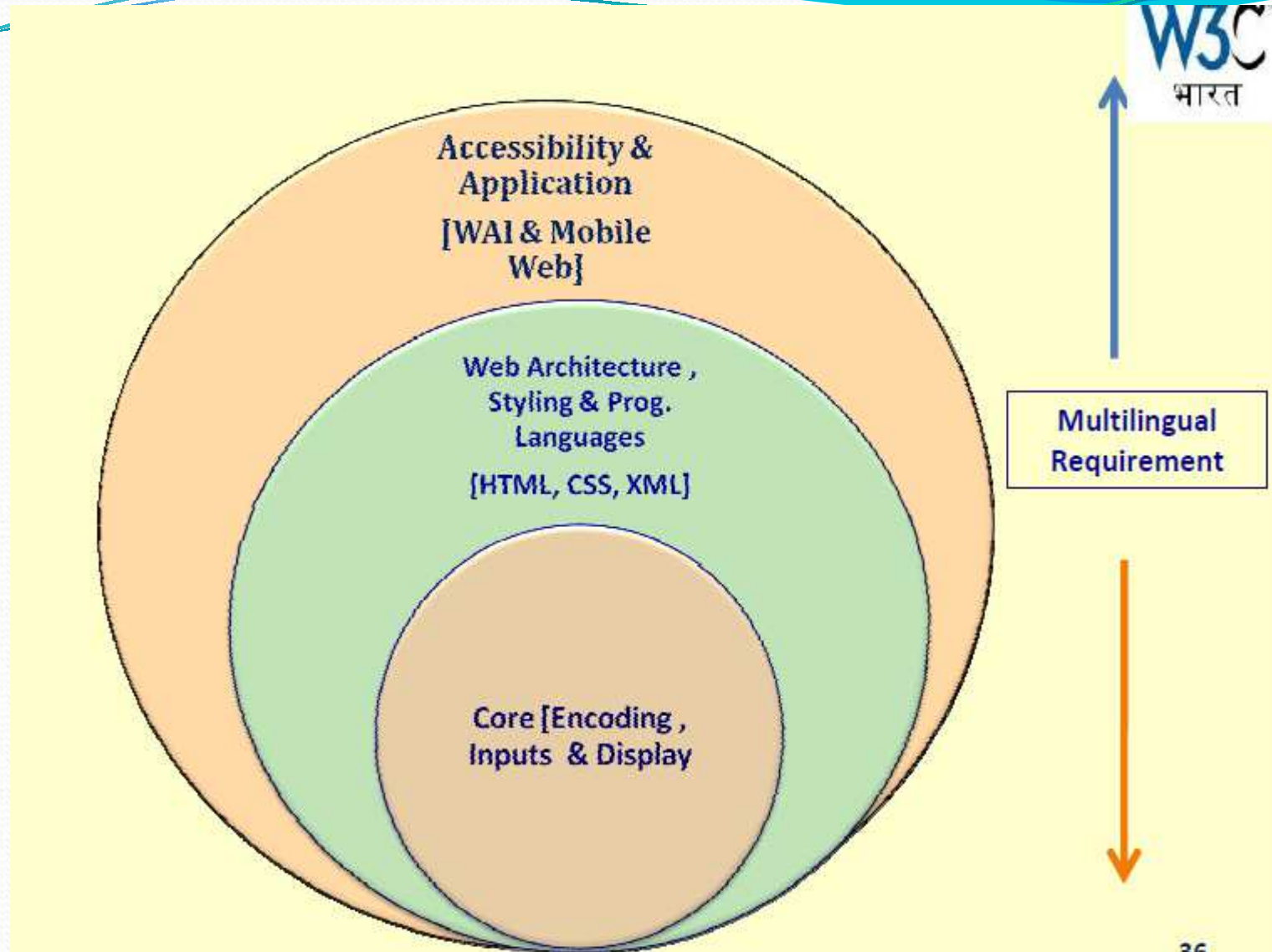
Major Scripts and Corresponding Languages in India



~ Courtesy: Swaran Lata , Country Manager , W3C India



Standardization Aspects for Supporting Multilingual Web in Indian Languages



Character Encoding : UNICODE

- Basis of Multilingual Web.
- All data exchange would be possible seamlessly across devices and Platforms
- Unicode Encoding for all 22 Constitutionally Recognized Indian Languages Complete.
- **Unicode declared as a Standard for Data Storage for Web Based E-Governance Services in India**

~ Courtesy: Swaran Lata , Country Manager , W3C India



Styling Issues in Indic Languages

Drop Letters in Indian languages

- Issues for Indian Languages with respect to first character used in Hindi, Malayalam, Bengali, Telugu and Gujarati etc.

क्रि केट के लिए मशहूर ब इसके लिए दीवानगी दिखाने वाले खेल प्रेमियों के देश भारत में अगर कोई अन्य खिलाड़ी प्रशंसा हासिल करता है तो यह किसी उपलब्धि से कम नहीं है। फिर हम जिस खिलाड़ी की बात कर रहे हैं उसने न केवल प्रशंसा अर्जित की है, बल्कि स्मरण भी बनाया है। लोग अब उन्हें उनके नाम से जानते हैं और यही अबुल अदवाल की क्रिकेट क्रेजी देश भारत में सबसे बड़ी उपलब्धि है। गोलफ कोर्स में तो अब्दुल अब्बल खिलाड़ी भी हैं ही। अब्दुल ने जब अमेरिकी पीजीए

বিজে পি-র সভাপতি রাজনাথ সিং জোরগলায় জানিয়ে দিয়েছেন, রামমন্দিরের প্রশ্নে বিজে পি-র আস্থা ও নিষ্ঠা কই মা কি লাল টলাতে পারবে না।

సో మహారం మార్కెట్ ఖరీదా రాభవించి. మార్కెట్లో ఉత్తీపన ప్యాకేజీ వార్తలు రావడంతో సెన్సిక్స్ పైపైకి పోయింది. సెన్సిక్స్ 483.03 పాయింట్లు రాభవడి 9583.89 వద్ద ముగిసింది. నిఫ్టీ 76.80 పాయింట్లు రాభవడి 2910.90 పాయింట్లు వద్ద ముగిసింది. రాభవడిన కంపెనీల పేర్లు ఖెల్, హెచ్ఐఎఫ్ఎస్, హెచ్ఐఎఫ్ఎస్ఇన్ఫర్, ఇస్కోస్, ఎల్ఎస్ఐ, మారుతి, రిలయన్స్, ఎస్ఐఐ, టాటామోటార్స్, టాటా పవర్, టీసీఎస్, షిప్రో, వోలెంటో ఇన్ఫర్, ఐటిసి, ఎన్టిపీసి, ఐఎస్ఐసి, మండ్రి కంపెనీలు రాభవడ్తాయి

ഇ ലേഖകനെ വളരെയധികം കർഷിച്ചിട്ടുള്ള വിഭാഗം വ്യക്തികൾ കണ്ണൂരിൽത്തന്നെ നാലുപേർ നാരായണമനോന്റെ ഈ വിഭാഗം വ്യവസ്ഥ വികൾ യുഗോവിന്റെ നോവൽ വിവർത്തനമായ 'പാവങ്ങൾ'

যুঁ তুঁতু কবিদ্বন্দ্ব জিস, বায়, তুঁতুতু ভারতমাং যৌক্সলানী যুঁতুতুতু মলবিত তারীখীনি গ্রন্থমা নির্দেশ আখ্যো অনে ভারতনা রাজসারথীখীমা ভুলমল মথী মথী, বাহলমমা অ যুঁতুতুতুতু মনাবার জাউতাত নভীতী, যত্নে ভুলে রাজসীয পঞ্জী অগ্ন মানবা আখ্যা চৈত যুঁতুতুতুতু তারীখীনি গমে ল্যাই জাউতাত মথী মথী চৈ. আখু বায় তৌ অ দিলসখী যুঁতুতুতুতু আবারসংহিতানী অমল শব্দ মথী জায়. শাসক মৌরানা পঞ্জীনে অনে আত্ম কথীনে সরসারনা বিবিধ মন্ত্রালয়ীনা মন্ত্রীখীঅ যুঁতুতুতুতু মথীলো আত্মনা আত্মনী অনেত যৌজনাখীনি জাউতাত কতবানী তৈমারী কথী

அன்று செவ்வாய்க்கிழமை. வழக்கத்தையிடச் சற்று முன்னதாகவே அகல்பா விழித்துக் கொண்டு விட்டாள்.

~ Courtesy: Swaran Lata , Country Manager , W3C India

Underlining of characters

- There is some examples of Indian languages in which Matra's are not readable due to underlining of characters

Hindi - अन्य भाषाओं में भी अनुवाद

Punjabi- ਰਾਜ

Bengali- তাই পুরোনো আর্কাইভ একটু ওলট পালটে।

Guajarati - સરદાર ગુજરી

Marathi- मराठी मुला मुलींची नावे

Tamil- நீரிற்குமிழி யிளமை நிறைசெல்வம்

Telugu - శ్రీలం ప్రజక్షు TV9 ప్రోగం " డ్యూమ్ ఇన్ డంజర్ " పార్ట్

~ Courtesy: Swaran Lata , Country Manager , W3C India

- Vertical arrangements

च ^ॐ	व	or	व	or	वक्	श	श
द	क्ता		क्		ता	क्ति	or
			ता				क्
							ति

- Bullets and Numbering

अ)	U+0906
आ)	U+0908
इ)	U+0907
ई)	U+0908
उ)	U+0909
ऊ)	U+090A
ए)	U+090F
ऐ)	U+0910
ओ)	U+0913
औ)	U+0914
...	

ॐ)	U+0A73
अ)	U+0A06
आ)	U+0A08
इ)	U+0A72
ई)	U+0A38
उ)	U+0A39
ऊ)	U+0A15
ए)	U+0A16
ऐ)	U+0A17
ओ)	U+0A18
औ)	U+0A19
...	

- Indentation of character



~ Courtesy: Swaran Lata , Country Manager , W3C India

Major Identified Problems in Styling :

- Grapheme Cluster Problems for Vertical Writing Style
- Drop Initial Views of the First Letter Element
- Bullets & Numbering issues
- Justification Problems
- Horizontal Letter Spacing Problems
- Most browsers are unaware of syllable boundaries for Indic scripts.

~ Courtesy: Swaran Lata , Country Manager , W3C India

Approach to be taken for Possible Solution

1.Grapheme Cluster Problems :

Adoption of UAX#29 (Unicode Text Segmentation Algorithm addressing the complexities of Indian Grapheme Clusters)

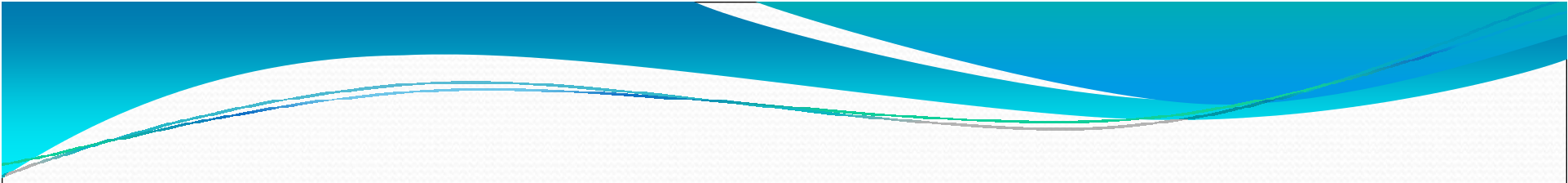
2.Bullets & Numbering issues

3.Justification Problems

4.Horizontal Letter Spacing Problems

- Development of Complete Mapping Table involving detailed requirements for document layout, typography and typesetting and calligraphic conventions , i.e. **Styling Manual**.
- The Rules developed in Styling Manual needs to be converted for inclusion in HTML and CSS

~ Courtesy: Swaran Lata , Country Manager , W3C India



Web Accessibility & Implementation

Issues for enabling Mobile Web in Indian languages

- **Character encoding**
- **Bandwidth and Cost**
- **Backward Compatibility with Legacy Devices**
- **Lack of standardization**
- **Fonts**
 - Bit map fonts (used by low cost handset)
 - True type fonts (used by high end handsets)
 - Open type fonts (currently in wider use)
- **Common Storage format**
- **Mobile messaging in indic languages**
- **Presentation Issues**

~ Courtesy: Swaran Lata , Country Manager , W3C India



Future Initiatives

Some of Future Initiatives:

- Multilingual Requirements for Indian Languages for HTML 5.0
- Multilingual requirements for Voice Browser in Indian Languages
- Gap Analysis for implementation of Mobile Web best practices in Indian Languages
- Multilingual Linked Data for E-Governance data

The Internationalization Best Practices Would be foundation of the all the above initiatives

~ Courtesy: Swaran Lata , Country Manager , W3C India



THANK YOU