

Semantics for Information Management in Agriculture

Embrapa's Experience

Rome/Italy, July 2nd-3rd 2015

*Ivo Pierozzi Júnior
Embrapa Agricultural Informatics
Campinas, SP, Brazil*



First words...

I think I´m not able to talk about “agricultural data interoperability in Brazil “ as it was suggested... I don´t have this whole scenario neither in my hands nor in my head...

I can give you a very fast overview about what Embrapa has been working last six years concerning basic semantic approaches aiming data and information interoperability.

I´m sure that there are other similar initiatives in Brazil, probably not only in Embrapa and not only for agricultural domain modelling, but they are incipient, fragmented and not yet implemented in a integrated basis or even not largely disseminated out of academic context.

So if we need solutions we have to develop/adapt it considering all kinds of idiosyncratic issues (technological, procedural, organizational, cultural, linguistic, epistemological, etc.)

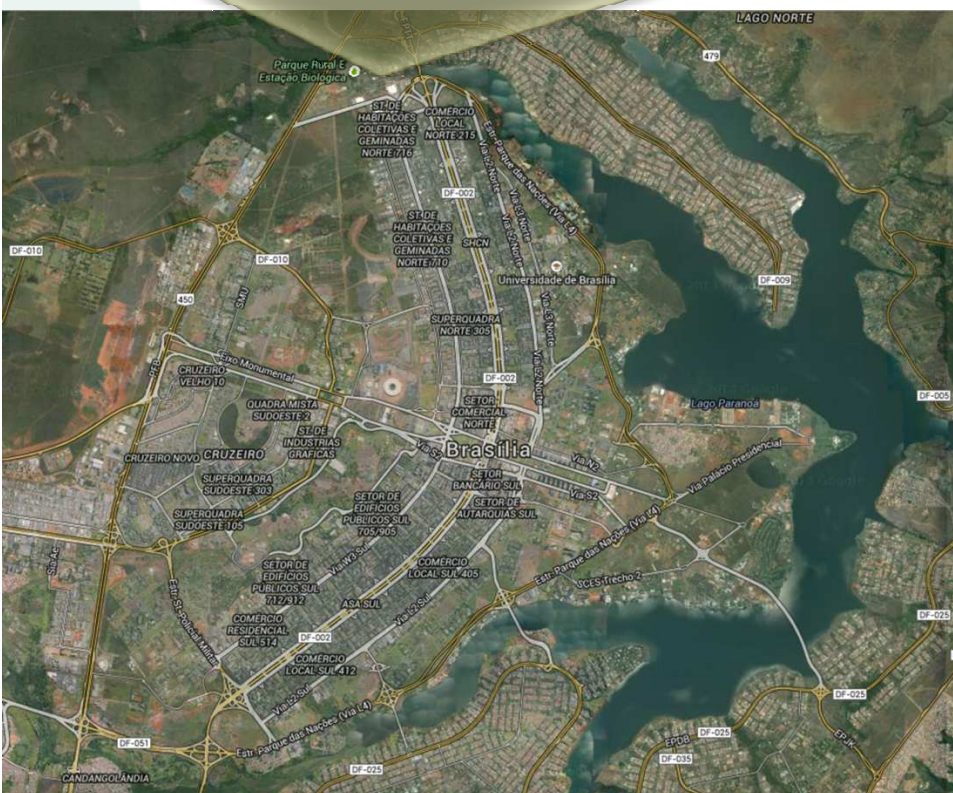


Embrapa

- PT: Empresa Brasileira de Pesquisa Agropecuária;
- EN: *Brazilian Agricultural Research Corporation*
- www.embrapa.br;
- linked and subordinated to the **Brazilian Ministry of Agriculture, Livestock and Food Supply**;
- 42 years old;

Geographical distribution:

- Head Quarter (15 central Units in Brasília, DF);
- 47 decentralized Units represented by thematic, eco-regional or products (e.g., cotton, rice and bean, beef cattle) research centers and services and distributed in all Brazilian states;
- 04 virtual laboratories abroad (Labex) in USA, Europe, China and South Korea;
- 03 international offices in Latin America and Africa;



People: 9,790 employees (2,444 researchers; 2,503 analysts; 1,780 technicians; 3,063 assistants);

Budget (2014): aprox. € 750,000,000 / USD 850,000,000 / R\$ 2.600.000.000.



Embrapa Agricultural Informatics

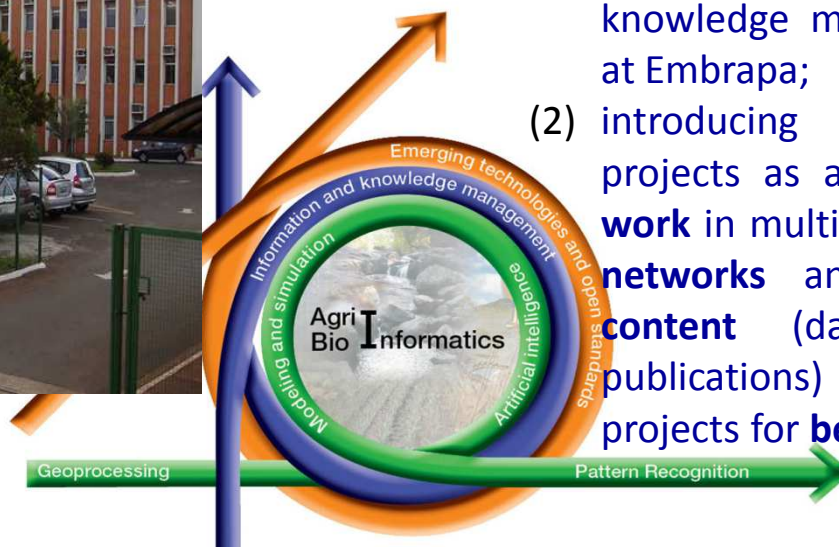


<https://www.embrapa.br/informatica-agropecuaria>

- thematic research center;
- Campinas, São Paulo State;
- computing and information technology applications for agriculture;
- About 100 people (60% RD&I; 20 Adm.; 15% TI; 15% Com; 10% TT);
- nine researcher groups;

Laboratory of Organization and Processing for Electronic Information

- knowledge organization and representation;
- “agriculturalKOS + NLP ” interfaces; (textual corpora, terminologies and conceptual structures), aiming:
 - (1) improvements of the general information and knowledge management corporate processes at Embrapa;
 - (2) introducing KOS/Terminologies in scientific projects as a facilitating tool of **collaborative work** in multi-inter-trans-disciplinary **knowledge networks** and preparing the **informational content** (databases, information systems, publications) produced during the execution of projects for **better interoperability** and **visibility**.



Working semantics in Embrapa...

Like answering 5W2H questions



Yesterday!!!
Last year!!!
Last decade!!!

OK!
What

OK!
Who

OK!
Where

When

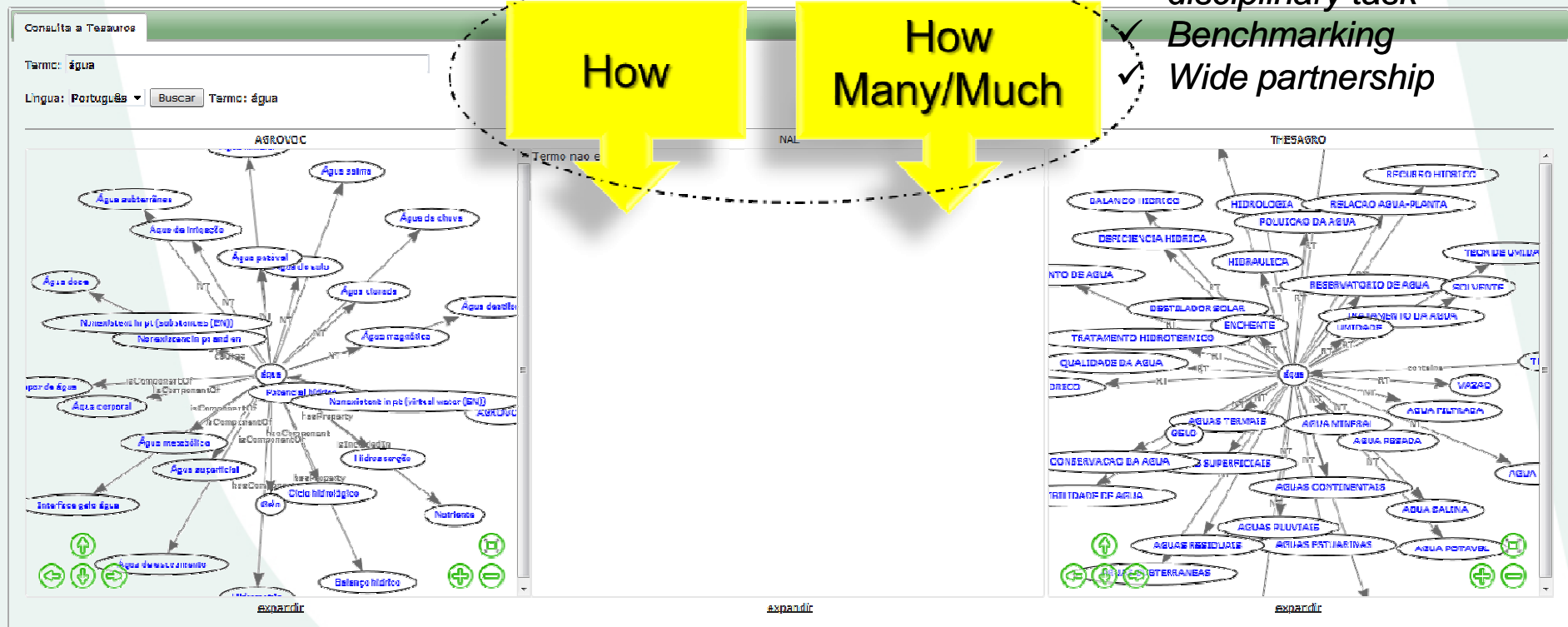
Why

Consulta a Tesouros Agrícolas

How

How
Many/Much

- ✓ Multi-inter-trans disciplinary task
- ✓ Benchmarking
- ✓ Wide partnership



Intentions, wishes, motivation ... :

Celerity and objectivity in pragmatic information retrieval and reuse to address complex problems

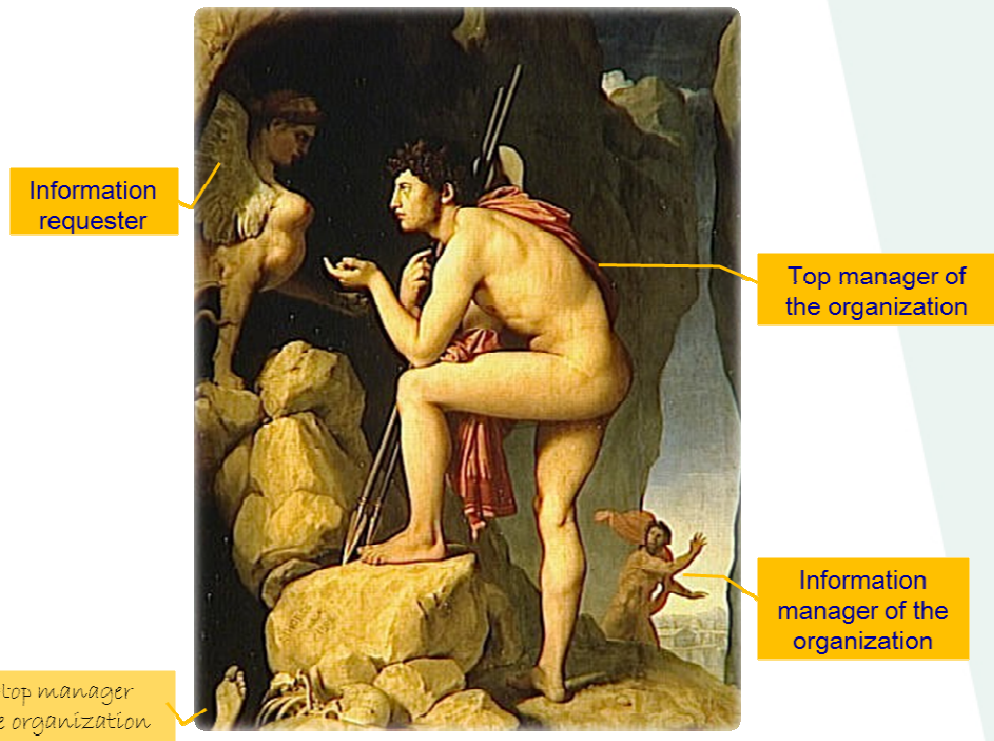


What would Embrapa **say** about this hypothetical information context?

Mitigation proposals to socioeconomic and environmental impacts due to silting caused by deforestation of riparian forest, with decreased water flow in the river bed, and due to the increase of population density and economic activities related to irrigated orcharding, and due to eutrophication caused by leaching of fertilizer (N) in Rio do Peixe in San Francisco river basin, Cerrado-Caatinga biomes transition after extreme drought events over the last five years (!!!)

Obs.:

1. If **to say = to have** (data, projects, publications, etc.), then: syntactic information recovery; Boolean rules; numerous lexical combinations to compose keywords for queries in databases and repositories... *Embrapa operates in this level!*
2. If **to say = to know** (reasoning), then: semantics = aggregated cognition value to data and information.
Embrapa needs technological and methodological evolution and innovation in this level.

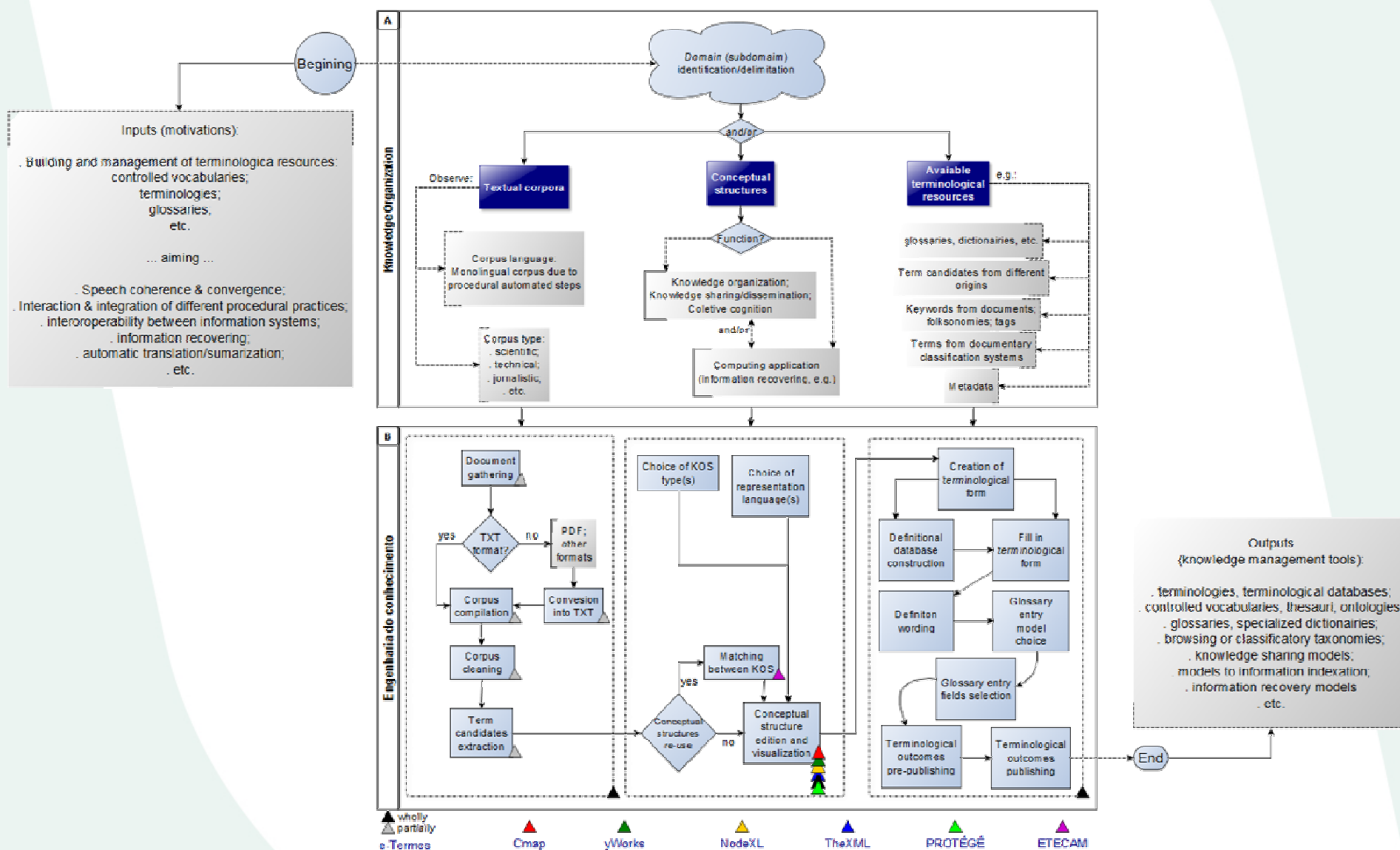


... Meanwhile ...

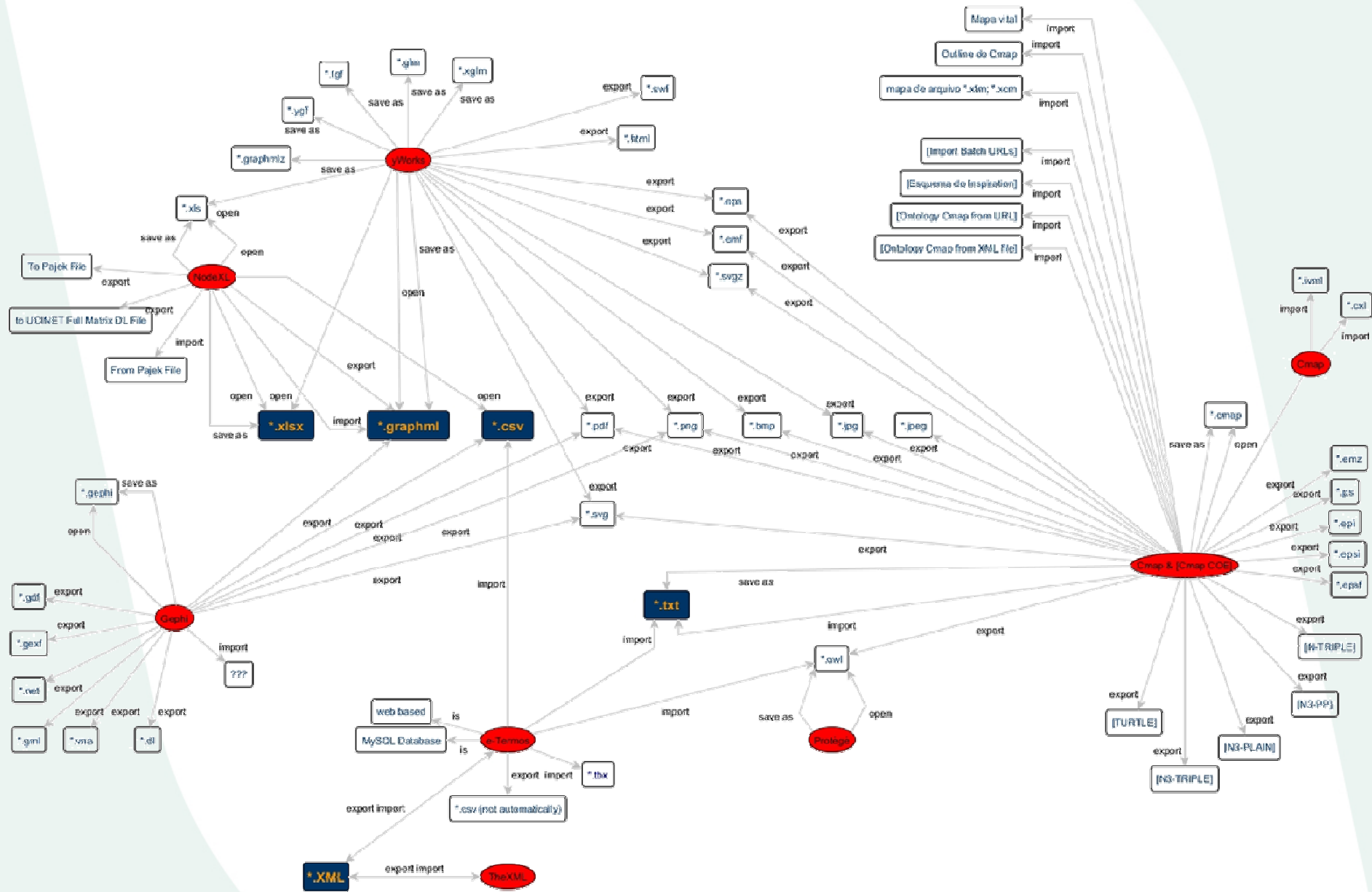
- ✓ Until recently (1-3 year ago) Embrapa had **no technological, methodological or procedural framework** to support the design, construction and management of a controlled vocabulary;
- ✓ Until now, **librarians just use controlled vocabularies to index documentary information** and they consult **manually** thesauri websites; until now, there was no automatized facilitations in this process.
- ✓ **We could not propose to Embrapa to build its own controlled vocabulary “starting from point zero”** (i.e., concept by concept, term by term, relationship by relationship);
- ✓ We start a project aiming **provide such framework and proposing to populate it initially with terminologies and conceptual structures re-used from other sources**;
- ✓ We design a very **broad proposal of knowledge organization and engineering**, based on **corpora linguistics, conceptual structures and terminologies**. In this process we suggest to use several software to run several steps;
- ✓ The need to re-use other sources has led to the development of a specific computational tool, an "automatic extractor of agricultural multi-lingual terms and conceptual structures" (**ETECAM** in Portuguese acronym) to automate the task of matching terms in thesauri. ETECAM was included as an alternative step in the proposed corporative framework mentioned above.

NOTE: ETECAM provides a “terminological” matching and based on comparison between the input term and its occurrence in the triples (concept-relationship-concept) registered in the thesauri; a real “semantic” matching could be achieved when Embrapa’s could implement technological evolution in this process.

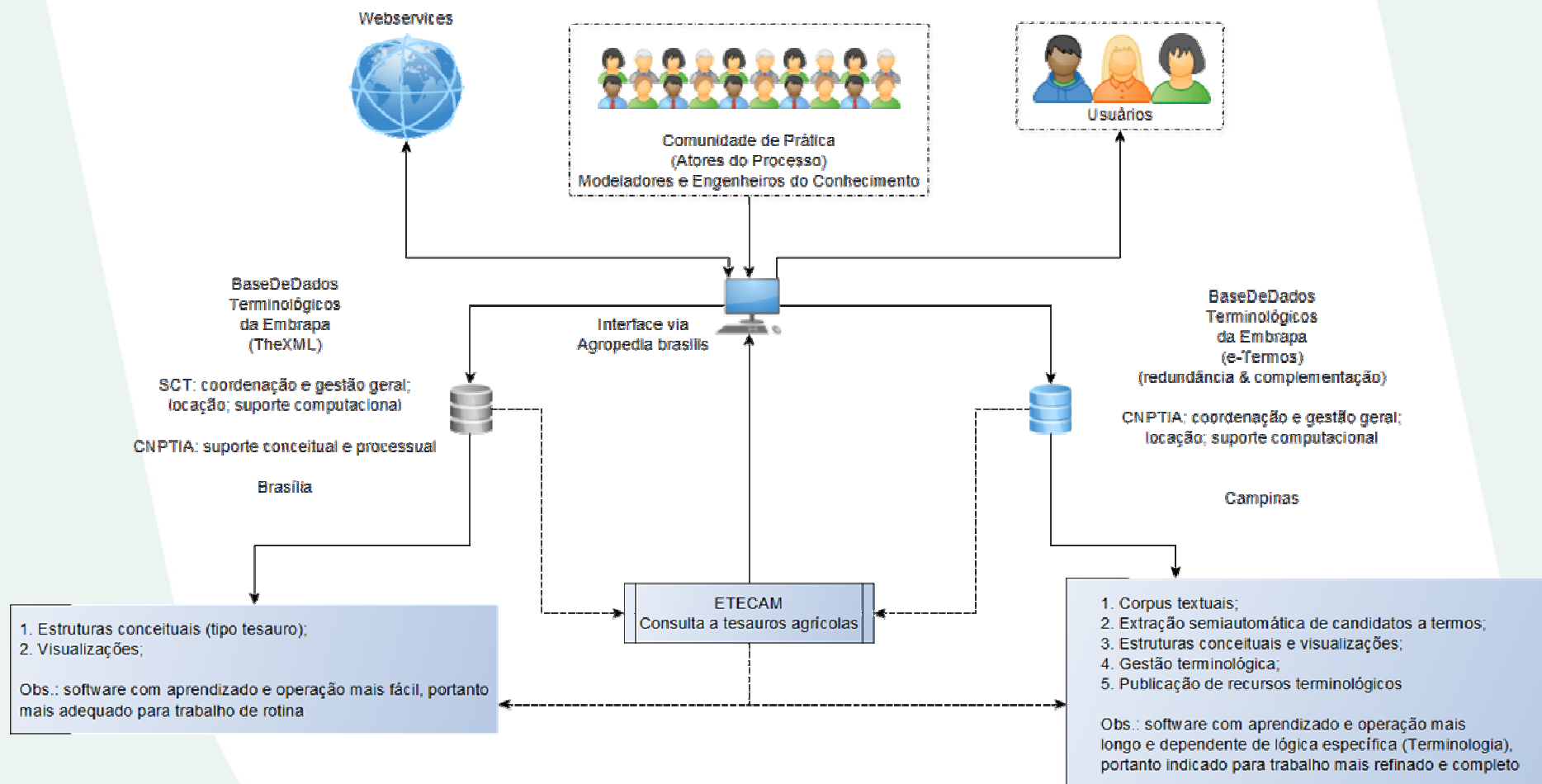
Embrapa's Knowledge Organization & Engineering Process

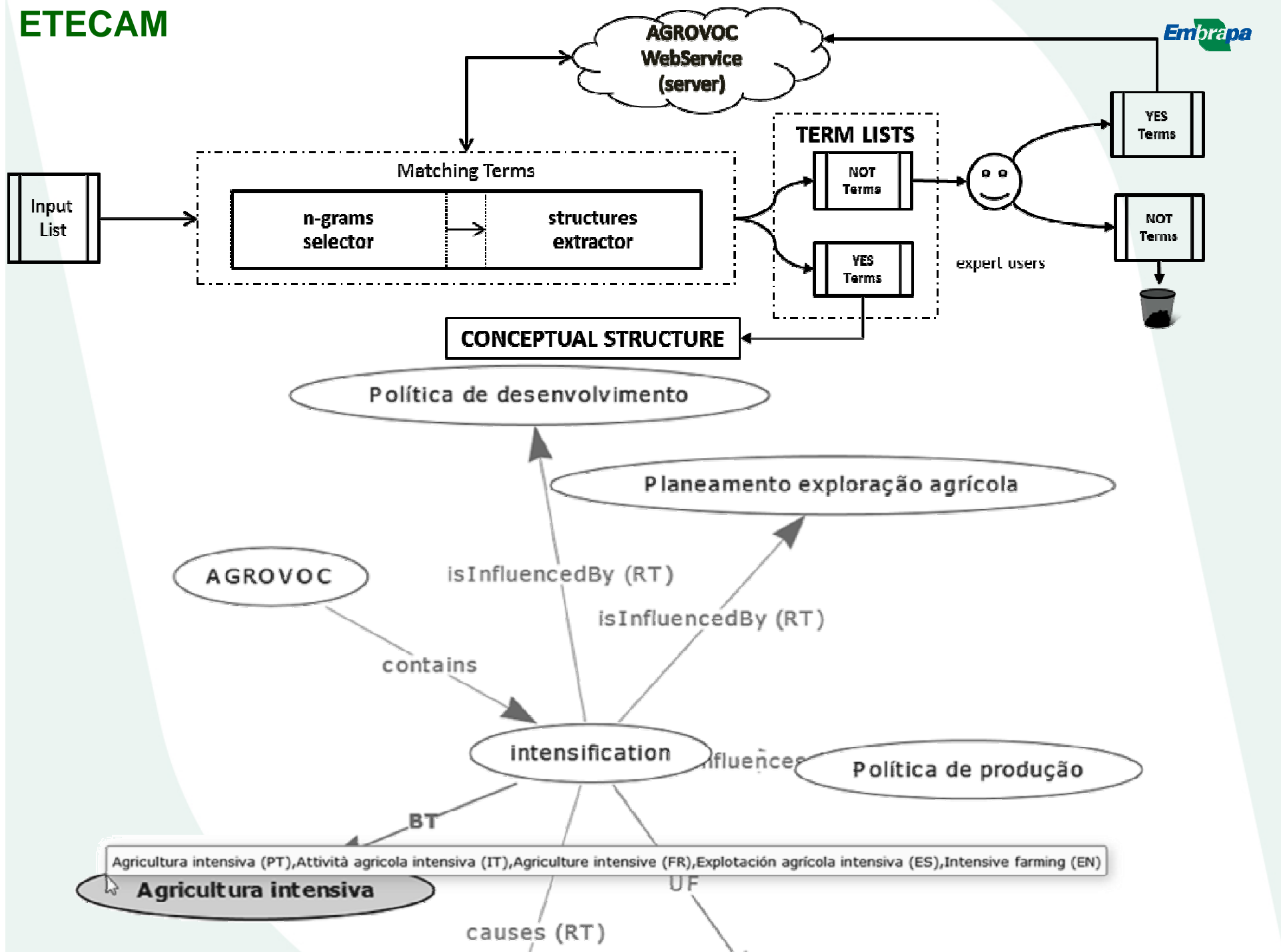


Software file formats for inter-changes ...



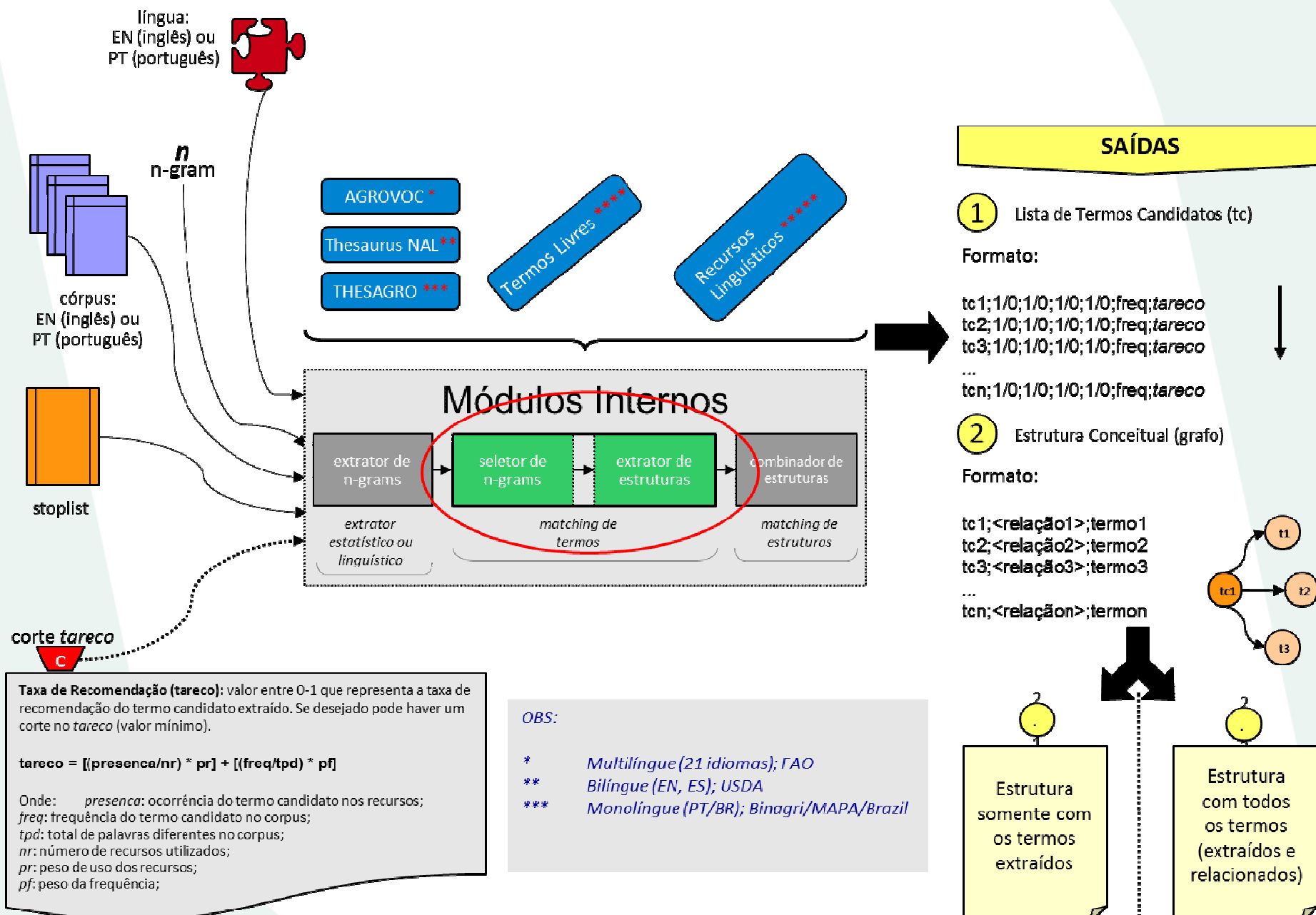
Embrapa's Framework for Knowledge Organization & Engineering





ETECAM

automatic extractor of terms and multi-lingual agricultural conceptual structures



Agrovoc, NAL Thesaurus and Thesagro conceptual structures visualization (side by side)

Consulta a Tesouros Agrícolas

Termo: água

Língua: Português

Buscar

Termo: água

AGROVOC

NAL

THESAGRO

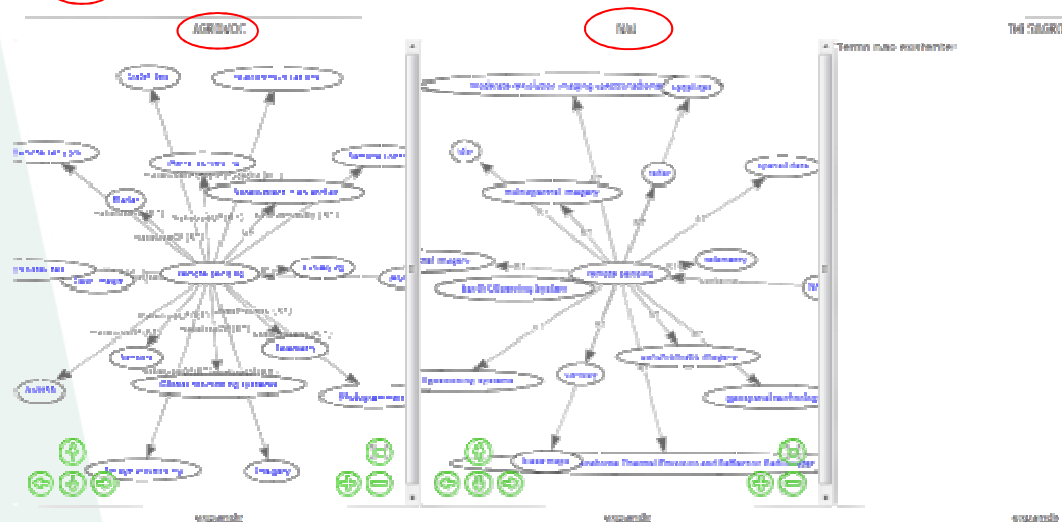
expandir

Embrapa Informática Agropecuária

Agrovoc, NAL Thesaurus and Thesagro conceptual structures visualization

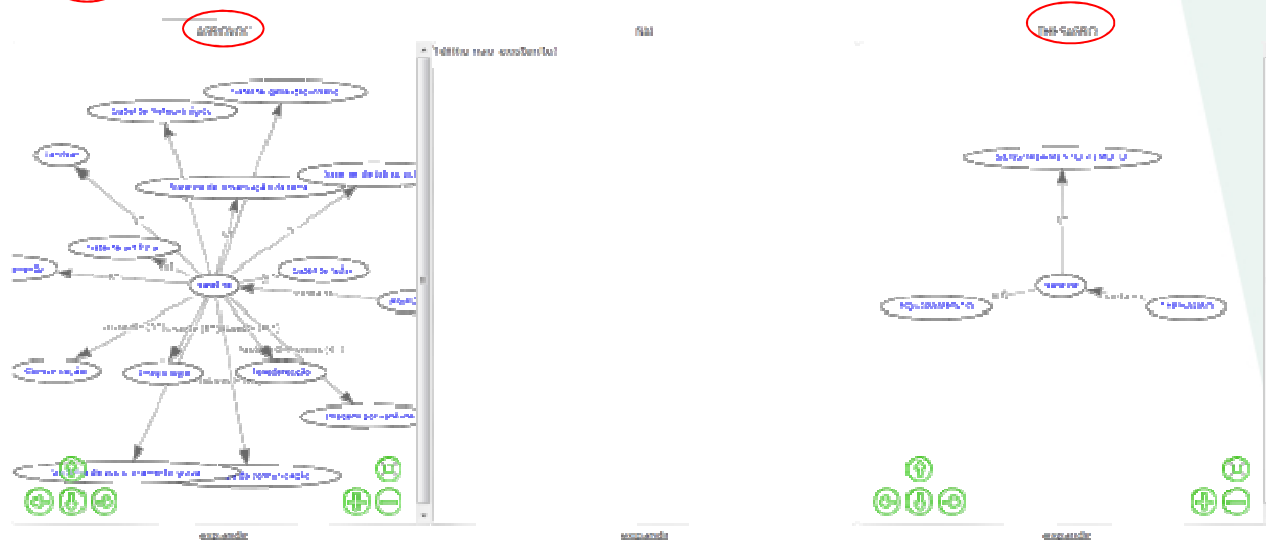
Termos: remote sensing

Lingua: **English** Buscar Termos: remote sensing



Termos: satélites

Lingua: **Português** Buscar Termos: satélites



PT/BR and PT/PT idiomatic differences



ENGLISH	PT/PT	PT/BR	DIFFERENCES DUE TO	OBSERVATIONS
Agricultural planning	Planeamento agrícola	Planejamento agrícola	Orthography	"planejamento" instead of "planeamento"
Reproduction control	Controlo da reprodução	Controle da reprodução	Orthography	"controle" instead of "controlo"
Sterile insect release	Libertação de insectos estéreis	Liberação de insetos estéreis	Orthography	"liberação" instead of "libertação"
Canopy	Coberto arbóreo ou arbustivo	Cobertura arbórea ou arbustiva	Orthography	"cobertura" instead of "coberto"
Land clearance	Rocadura	Roçada	Orthography	PT/PT and PT/BR must utilize the character "ç"; there is no meaning for the word written with character "c"
agricultural research	Investigação agrária	Pesquisa agrícola	Other term used in PT/BR	-
Poultry farming	Criação de aves de capoeira	avicultura	Other term used in PT/BR	-
bare fallow	Pousio inculto	Pousio de/com solo nu; pousio nu	Other term used in PT/BR	-
Tillage	Mobilização do solo	Trabalho do solo	Other term used in PT/BR	-
Farm management	Administração exploração agrícola	Gestão agrícola	Other term used in PT/BR	-
Irrigation rates	Dotação de irrigação	Taxa/grau de irrigação	Other term used in PT/BR	-
Sole cropping	Cultivo estreme	Cultivo puro	Other term used in PT/BR	-
Dung	Bosta	Esterco; estrume	Other term used in PT/BR	-
Plot size	Dimensão do talhão	Dimensão da parcela	Other term used in PT/BR	-
Fish processing	Tratamento do pescado	Processamento de pescado	Other term used in PT/BR	-
Food shortages	Penúria alimentar	Escassez alimentar	Other term used in PT/BR	-
Conservation tillage	Mobilização solo para conservação	Práticas de conservação do solo	Other term used in PT/BR	-
Biological control agents	Agente de luta biológica	Agente de controle biológico	Other term used in PT/BR	"controle" instead of "luta"
Weeding	Monda	Capina	Other term used in PT/BR	-
bumble bees	Abelhão	Mamangava; mamangaba	Special case: term from regional linguistic	Brazilian indigenous term
burn tillage *	-	coivara	Special case: term from regional linguistic	Brazilian indigenous term

PT/BR and PT/PT idiomatic differences



Arquivo Editar Exibir Histórico Favoritos Ferramentas Ajuda

AGROVOC Search

https://www.terminos.cnptia.embrapa.br/agrovocsearch/#

Pesquisar

Consulta de Termos

Consulta de Termos em Ordem Alfabética

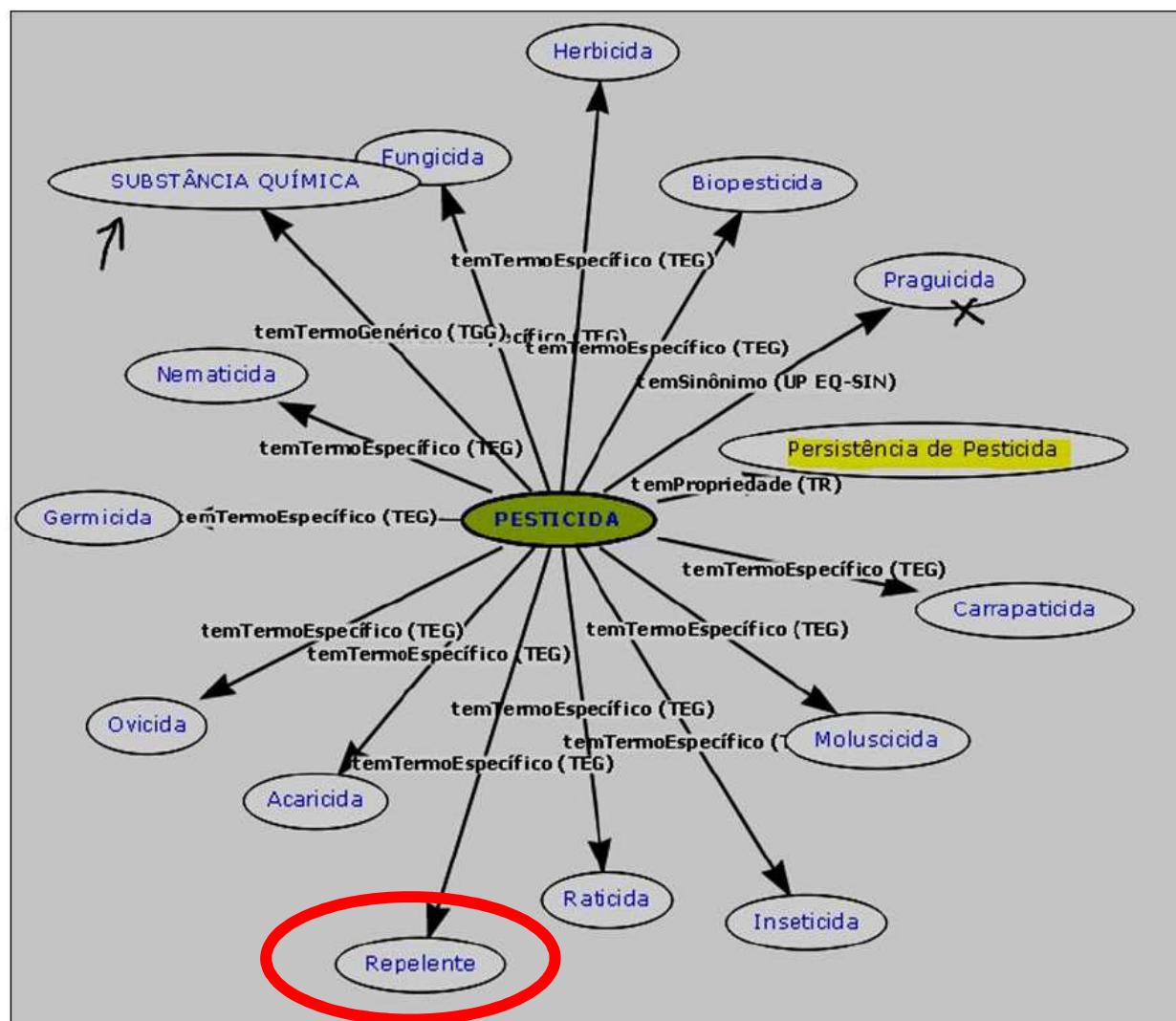
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Quantidade Total de Termos: 2085

Inglês	Português (PT)	Espanhol (ES)	Francês (FR)	Italiano (IT)	Português (PT-Br)
B (symbol)	B (símbolo)	B (símbolo)	B (symbole)	B (simbolo)	B (símbolo)
B-NINF					+
B-cells	Célula-b	Células-b	Cellula-b	Cellule B	+
B-lymphocytes	Linfócito-b	Linfocitos-b	Lymphocyte-b	Linfociti B	+
BA	6-benziladenina	6-benziladenina	6-benzyladenina	BA	+
BDV	BVD	BDV	BDV	BDV	+
BHA	BHA	BHA	BHA	BHA	+
BHT	Bht	Bht	Bht	BHT	+
BNF	FBA	FBN	FBA		+
BOD (biochemical oxygen demand)	Necessidade bioquímica de oxigênio	DBO (demanda bioquímica de oxígeno)	DBO (demande biochimique d'oxygène)	BOD (domande biochimica di ossigeno)	DBO, DBOS (demanda bioquímica de oxigênio, demanda bioquímica do oxigênio)
BSE (bovine spongiform encephalopathy)	EEB (encefalopatia espongiforme bovina)	EEB (encefalopatia espongiforme bovina)	ESB (encéphalopathie spongiforme bovine)	BSE (encefalopatia spongiforme bovina)	+
BYDV	Bydv	Bydv	Bydv	BYDV	+
Ba (symbol)	Ba (símbolo)	Ba (símbolo)	Ba (symbole)	Ba (simbolo)	+
Babacu palms				Palme babacu	+

Thesagro re-engineering⁽¹⁾: conceptual structures visualization

FIGURA 27 – Relação gênero-espécie: conceito PESTICIDA



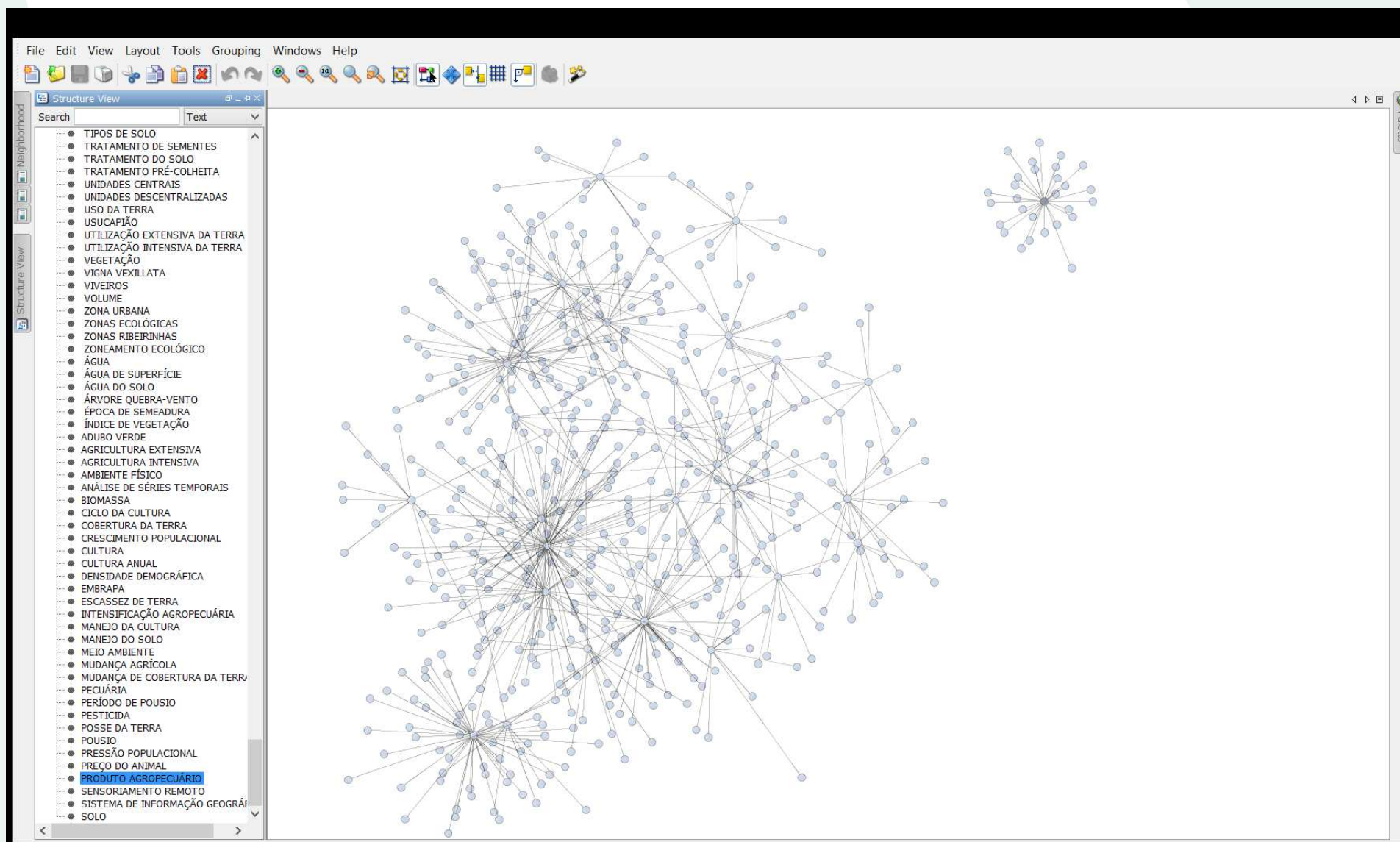
PESTICIDA
 UF PRAGUICIDA
 NT ACARICIDA
 NT CARRAPATICIDA
 NT FUNGICIDA
 NT GERMICIDA
 NT INSETICIDA
 NT MOLUSCICIDA
 NT NEMATICIDA
 NT PERSISTÊNCIA DE PESTICIDA
 NT RATICIDA
 NT REPELENTE

Fonte: elaborado pela autora desta tese (tela capturada do sistema e-Termos).

(1)

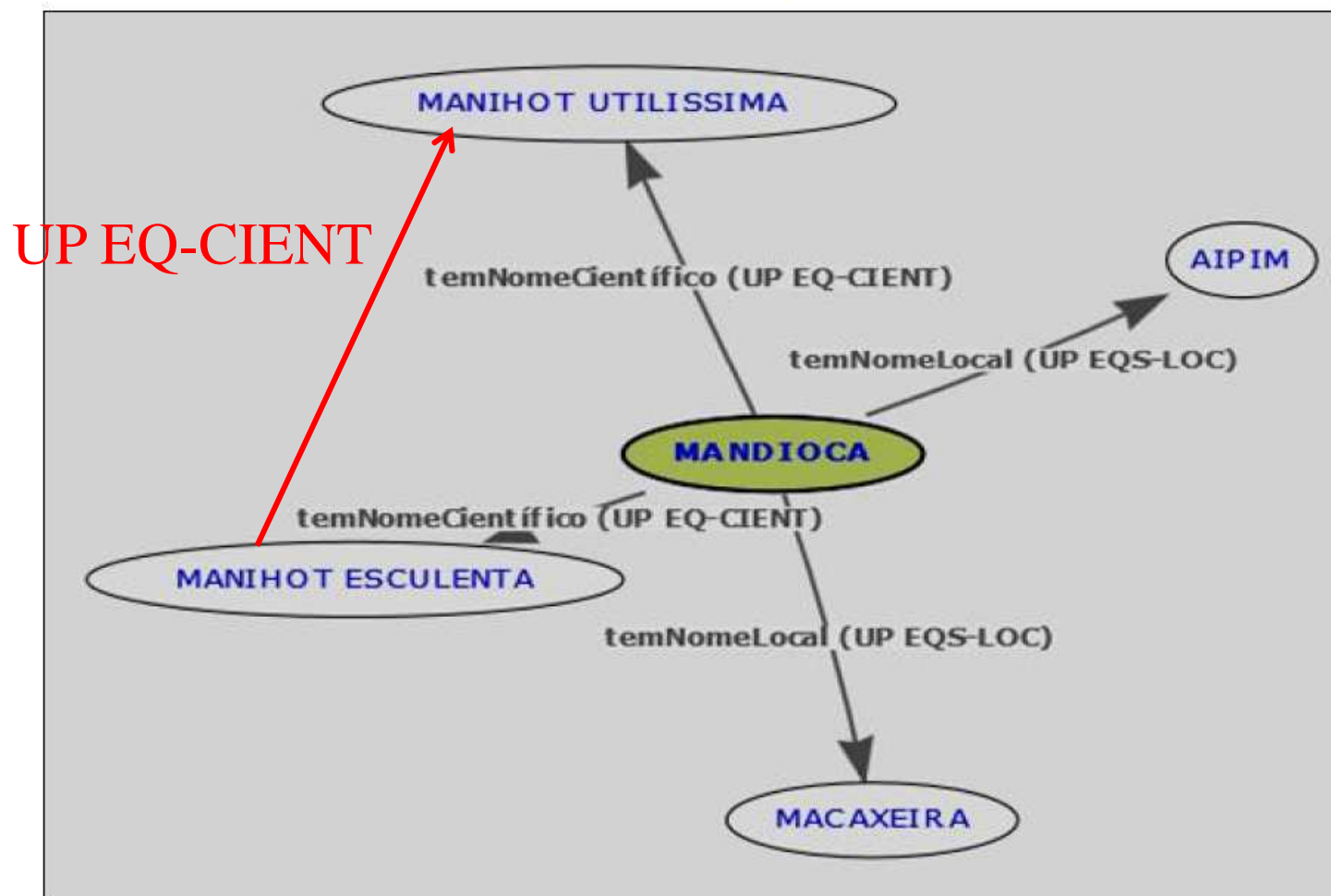
- SOERGEL, D. et al. Reengineering thesauri for new applications: the AGROVOC example. Journal of Digital Information, v.4, n.4, 2004;
- LAUSER, B. et al. From AGROVOC to the Agricultural Ontology Service: Concept Server an OWL model for creating ontologies in the agricultural domain. In: INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATIONS, 2006, Colima, Mexico. Proceedings... México: DCMI, 2006.

Thesagro re-engineering: conceptual structures visualization



Thesagro re-engineering: conceptual structures visualization

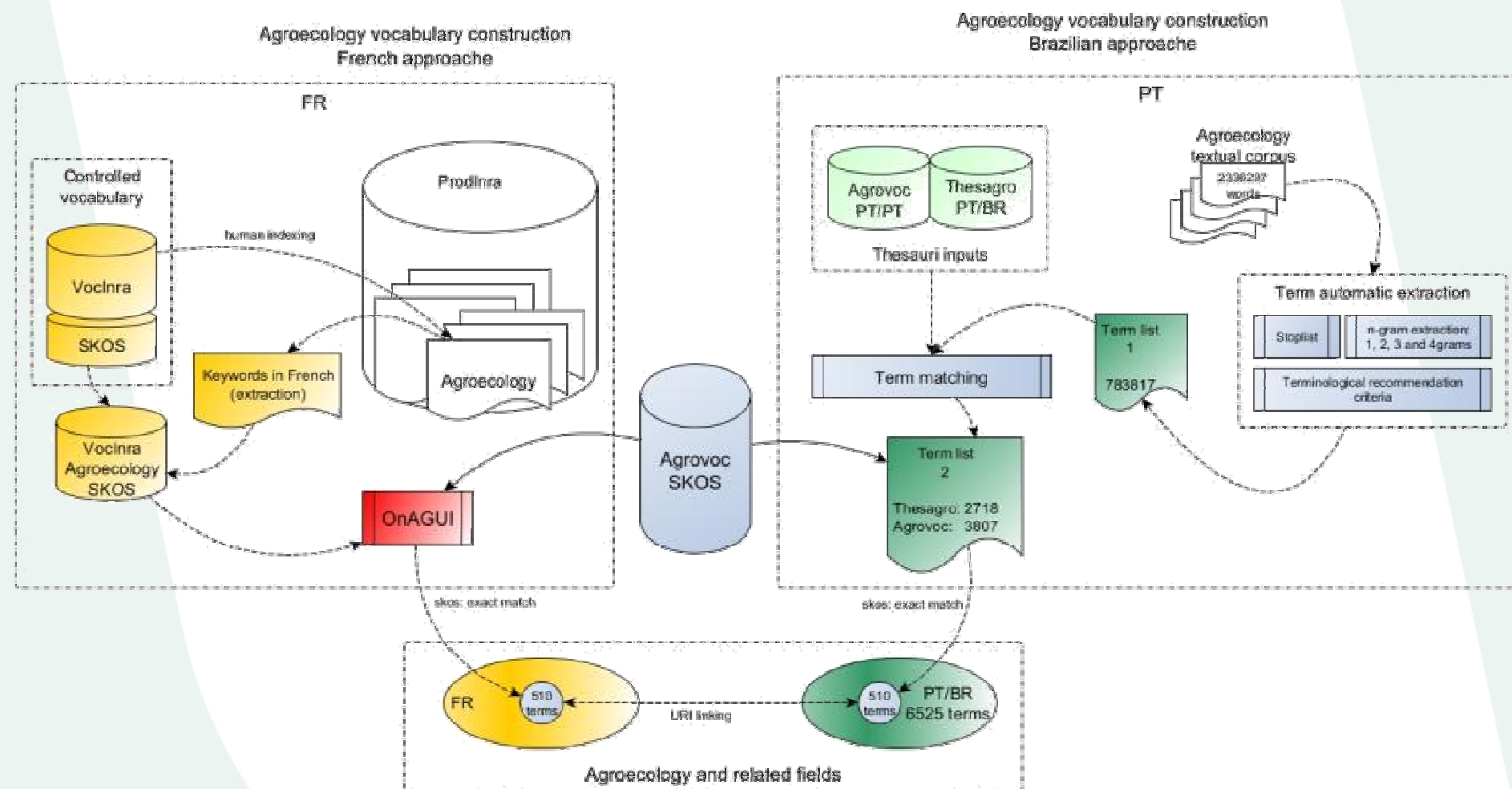
FIGURA 25 – Representação de MANDIOCA, a partir modelo de Soergel (2004; 2006)



Fonte: elaborado pela autora desta tese (tela capturada do sistema e-Termos).

Inter-linking INRA and Embrapa vocabularies (Agroecology)

Methodological Approaches



Thank you !
Merci beaucoup !
Muchas gracias !
Grazie molte !
vielen Dank !
बहुत बहुत धन्यवाद
Muito obrigado !

ivo.pierozzi@embrapa.br
Skype: ivo.pierozzi.jr