

On the rise of Machine Learning for scientific text mining

Patrice Lopez

LISA VIII - Strasbourg - June 8, 2017

Machine learning for scientific text mining

- Machine Learning approaches surpass today rule-based/human handcrafted approaches for all tasks
- Machine Learning has already a concrete impact on users of scientific information, in particular for:
 - ➔ assisting users
 - ➔ large-scale tasks

Large-scale Scientific Text and Data Mining (TDM)

- The core issue of TDM is automatic extraction of knowledge from scientific data
 - ➔ from scientific documents in particular for text mining
- TDM aims at answering research questions, thus addressing the largest possible relevant corpus
 - ➔ issue of scaling: we don't have reliable metadata, we don't have reliable content (PDF), we don't have homogeneous data schema, etc.

from PDF to TEI ?



Cow (structured data)



Hamburger
(unstructured data)

“Converting PDF to XML is a bit like converting hamburgers into cows.”

Michael Kay (<http://lists.xml.org/archives/xml-dev/200607/msg00509.html>)

Inspired from: Duncan Hull

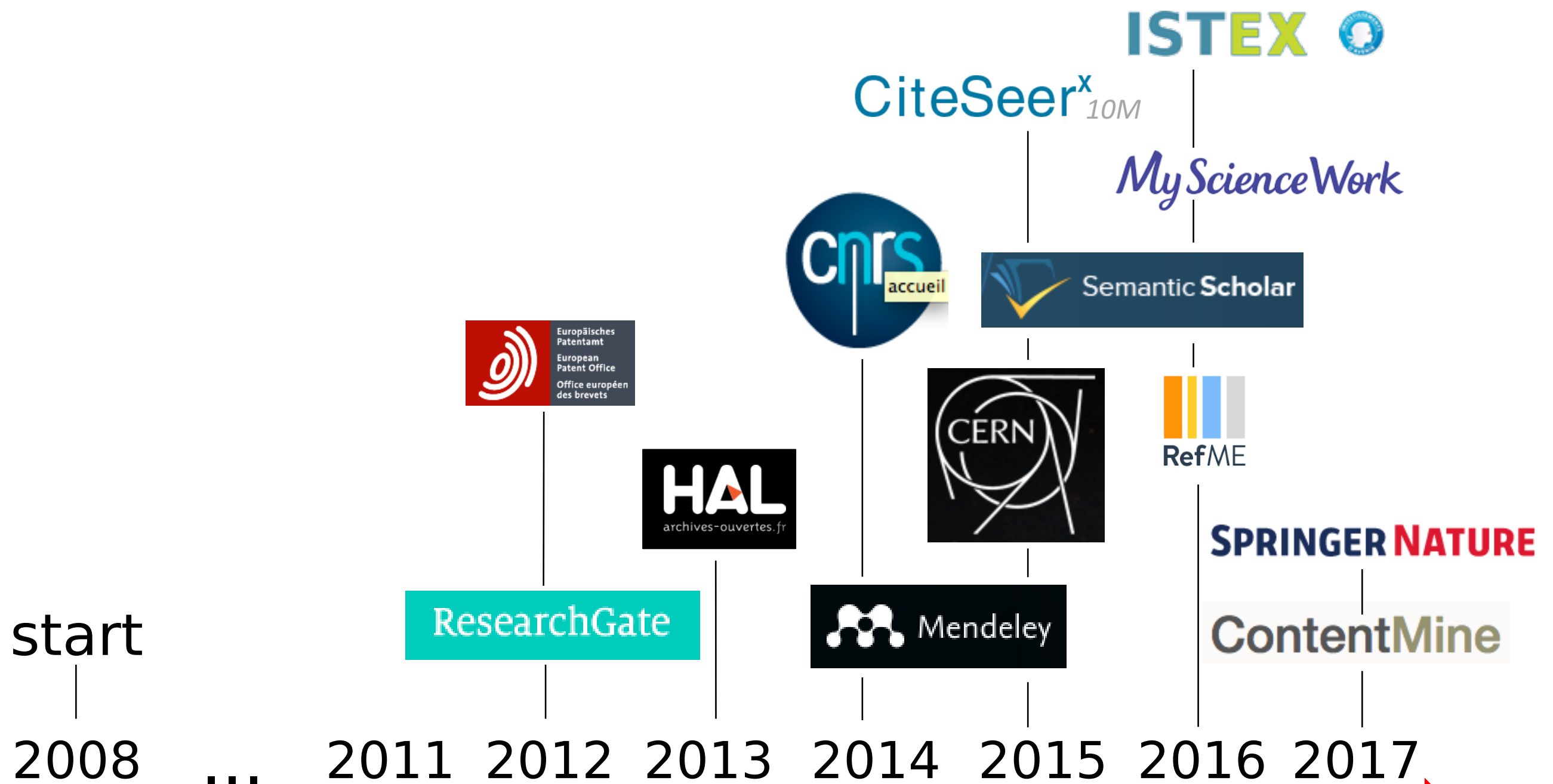
GROBID = automatic structuring of PDF documents

- Machine learning
- Open source Apache 2

<http://grobid.science-miner.com>

Accuracy and scaling

- Evaluation on 26 000 PDF, 92% header metadata extraction, 83.9% perfect citable data - Mendeley (2015)
- ISTEEX (18M publisher articles):
 - ➔ 75% perfectly recognized bibliographical references, up to 90% for article after 2000
 - ➔ 1 million PDF processed en 24h (Xeon 10 CPU, 10 GB memory, 3GB used in average, 9 threads) - 11,5 PDF/s



See all ›
360 Reads

See all ›
31 Citations

[Add supplementary resources](#) |

"...one [73]. Besides this, Lipinski et al. [96] compared many header parsers to conclude that GROBID [95] is the best parser. The study was conducted on arxiv dataset and can be tested for a large dataset. ..."

[Expand Text](#) [See in context](#)

A survey on scholarly data: From big data perspective

 Full-text available · Article · Jul 2017 · Information Processing & Management

 Samiya Khan  Xiufeng Liu  Kashish Ara Shakil  Mansaf Alam

[Read full-text](#)


2 Comments 0 Citations

"...[8]and potentially provides advantages over a word-based approach similar to the CRF model of Lopez[5]. In addition to the reduced complexity of the model, a linebased approach can capture patterns that..."

[Expand Text](#) [See in context](#)

Reference String Extraction Using Line-Based Conditional Random Fields

 Full-text available · Article · May 2017

 Martin Körner

[Read full-text](#)

0 Comments 0 Citations




researchgate.net

Inbox (1693) ▾ Drafts (217) ▾ Sent ▾ Notes ▾ Flagged Save

From	Subject	Date Received	S..
ResearchGate	Patrice, we found one more citation...	May 30, 2017 10:01	25 K
ResearchGate	Patrice, this job could interest you	May 29, 2017 09:47	29 K
ResearchGate	Congratulations Patrice, you reac...	May 24, 2017 10:41	24 K
ResearchGate	Patrice, this job could interest you	May 22, 2017 09:46	30 K

ResearchGate



Patrice, we found one more citation of your research
GROBID: Combining Automatic Bibliographic Data Recognition a...

[View citing research](#)

<https://www.semanticscholar.org/>

Semantic Scholar

SUBSCRIBE SHARE SIGN IN ?

Cut through the clutter.

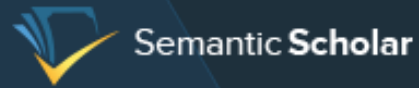
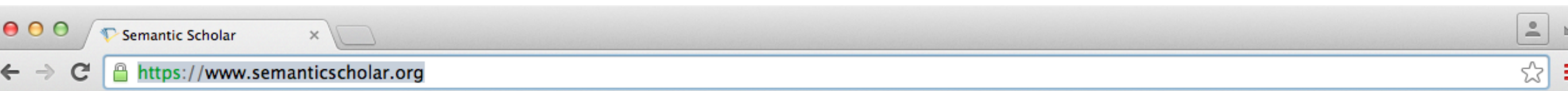
Home in on key publications, citations, and results.

Search

Try: [Christof Koch](#) [Songbird Basal Ganglia](#) [Deep Learning](#)

Our corpus currently includes only computer science and neuroscience papers.

<https://www.semanticscholar.org/>



[SUBSCRIBE](#) [SHARE](#) [SIGN IN](#) [?](#)

Annotating Temporal and Event Quantification

Harry Bunt, James Pustejovsky • Proceedings of 5th ISA Workshop • 2010

≡ 4 Excerpts

Language resource management
Iso-Dis • Language resource management

"But as has been pointed out, there is no clear way to embed an event within a temporal quantifier expression (Pratt-Hartmann, 2007, *Bunt and Pustejovsky, 2010*)."

Language resource management
Iso-Dis • Language resource management

"*Bunt and Pustejovsky (2010)* explore the option of explicitly marking the distributive property (Bunt, 1985) of the quantification in the annotation directly."

Language resource management
Iso-Dis • Language resource management

"A temporal expression such as 3 hours is expressed as a TIMEX3 of type DURATION, with the interpretation of a "time amount" (*Bunt and Pustejovsky, 2010*)."

TEI P5:-Guidelines for Electronic Texts
Lou Burnard, Syd Bauman • TEI P5

"The details of this proposal are more fully presented in *Bunt and Pustejovsky (2010)*."

The Language of Time: A Reader

Patrick Blackburn • Computational Linguistics • 2006

Abstract & Details

References

Citations

Statistics

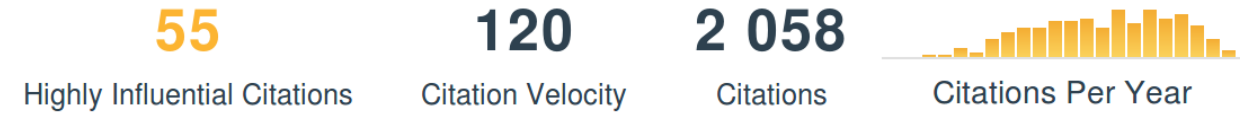
Related Publications

Our corpus currently includes only computer science and neuroscience papers.

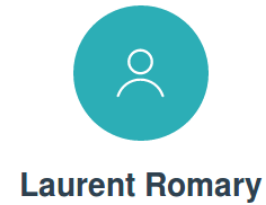
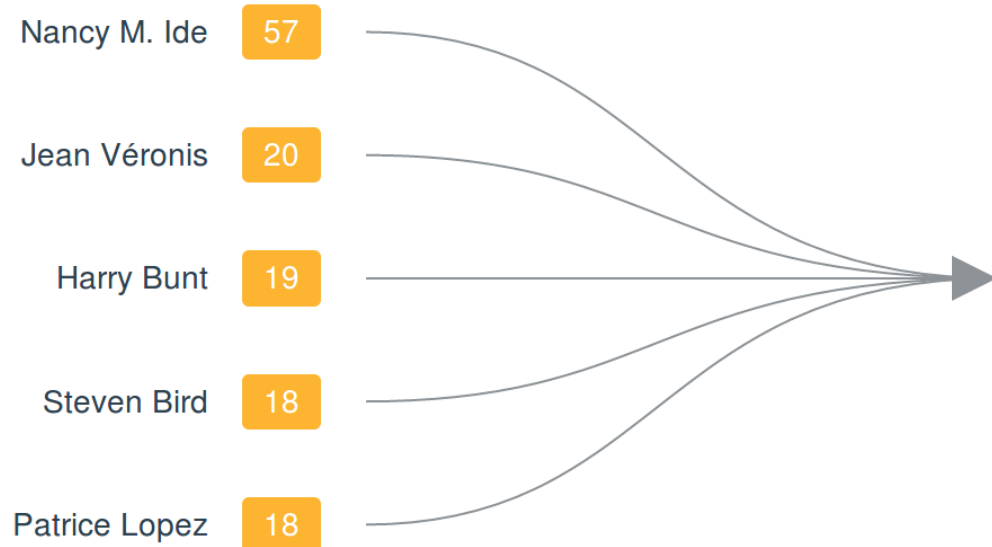


Laurent Romary

[Share](#) [Suggest Changes](#)

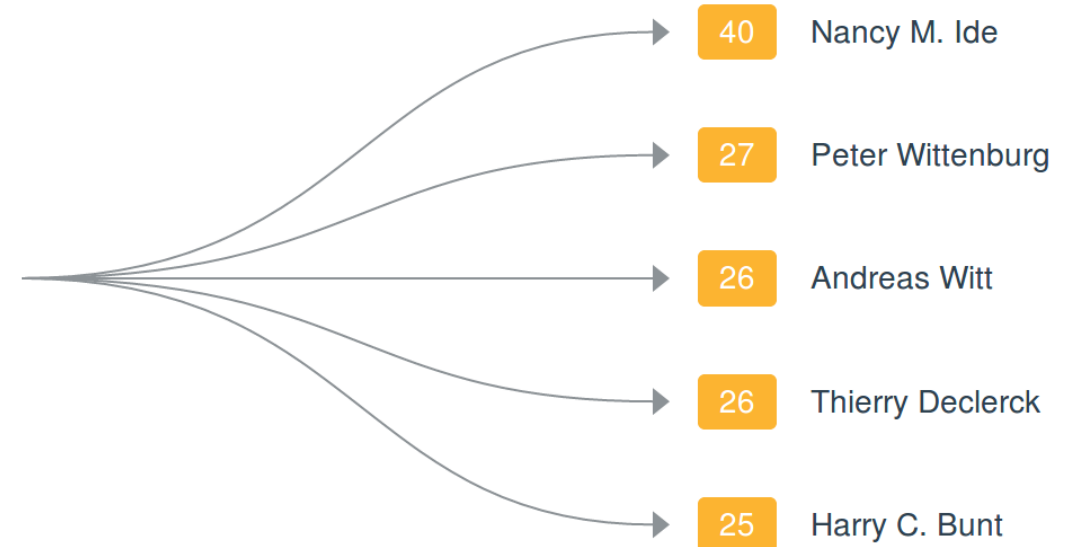


Authors who most influenced Laurent Romary:



Laurent Romary

Authors most influenced by Laurent Romary:







A Generic Formalism for Encoding Standoff Annotations in TEI

Javier Pose Rodriguez
European Patent Office
jposem@epo.org

Patrice Lopez
INRIA Saclay
patrice.lopez@inria.fr

Laurent Romary
INRIA Saclay
laurent.romary@inria.fr

Abstract

This article describes a proposal for a comprehensive and consistent encoding of standoff annotations in the frame of the TEI standard. The encoding requires the extension of the current TEI schema with three additional elements providing a generic and flexible structure for encoding standoff annotations in multiple layers or levels of annotations. The proposal, part of a bigger project still in progress, is implemented and validated for an heterogeneous set of annotations in a large scale collection of patent documents.

Keywords: textual resources, TEI, annotations
Word count: 1978 (without references and figures)

1. Motivation

In the field of digital document representation, various proposals for the encoding of annotations in relation to an existing XML document have been proposed. When encoding such annotations, two different approaches can be considered: inline annotations, where the markup elements of the annotations are directly added into the annotated data, and standoff annotations, where the markup elements of the annotations reside in a location different from the location of the data being described.

Whereas the inline annotation has as main advantage its simplicity and easy maintenance, it also presents numerous drawbacks when implementing multiple types of annotations on the same source text, a common requirement in document engineering. In such cases, the standoff annotation mechanism has shown to be a better option (Dipper 2005).

The standoff annotation mechanism allows the definition of new layers of information on the top of the source textual data without disturbing existing ones, and are therefore appropriate for encoding multiple concurrent annotations on a source text.

The Text Encoding Initiative (TEI) is the major XML standard for encoding and documenting primary data (Przepiórkowski and Banski, 2009). However, most of the existing text encoding standards that allow the encoding of annotations, and in particular standoff annotations, have been proposed outside the TEI: the standards developed in the frame of ISO TC37/SC4 (Idé and Romary 2006), XCES (Idé, Bohomme and Romary 2000), TIGER/XML (Mengel and Lezius, 2000), PAULA (Dipper 2005) or XStandoff (Stührenberg and Jetka 2009).

There have been a few attempts to encode standoff annotations within the TEI (Boot 2009, Przepiórkowski 2009) (Banski and Przepiórkowski 2009) (Banski and Przepiórkowski 2010), or the TEIWiki page for "Stand-off

use cases"¹. However, none of them offers a general framework for standoff annotations able to cope in a consistent way with a vast range of annotation types and to encompass the encoding standard proposed outside the TEI. These attempts only rely on specific combinations of already existing TEI elements for encoding the annotations in the context of specific projects.

The TEI guidelines disclose the basic referencing mechanisms for linking pieces of the source document with the annotations (see the TEI Guidelines: Chapter 16.9, 17.4 and 20.4), but they do not cover in a consistent manner the standoff annotations.

Our proposal aims at defining an extension of the TEI for representing standoff annotations fulfilling a set of comprehensive requirements (section 2). This supposes the introduction of additional elements described in section 3, allowing grouping and inter-operability with existing third party content annotation schema (section 4). The realization of the proposal is illustrated by heterogeneous automatic and manual annotations realized on a large scale collection of patent documents (section 5).

2. General encoding requirements

Our primary ambition is to provide an annotation framework going beyond the purpose of particular types of annotation, e.g. beyond linguistic annotation only, and beyond a particular family of documents, for instance texts only. We thus introduce the following requirements:

- **Expressiveness:** the framework must be able to encode heterogeneous types of standoff annotation models: multiple annotation layers, covering different content structures, references and segmentations.
- **Extensibility:** the framework is not limited to the encoding of standoff annotations of textual information, but support any type of source data, including audio, video, image, etc.
- **Uniformity:** the framework must provide a general mechanism that, even allowing different ways of encoding the annotation information, provides a

¹ http://wiki.tei-c.org/index.php/Stand-off_use_cases



A Generic Formalism for Encoding Standoff Annotations in TEI

Javier Pose Rodriguez
European Patent Office
jposem@epo.org

Patrice Lopez
INRIA Saclay
patrice.lopez@inria.fr

Laurent Romary
INRIA Saclay
laurent.romary@inria.fr

Abstract

This article describes a proposal for a comprehensive and consistent encoding of standoff annotations in the frame of the TEI standard. The encoding requires the extension of the current TEI schema with three additional elements providing a generic and flexible structure for encoding standoff annotations in multiple layers or levels of annotations. The proposal, part of a bigger project still in progress, is implemented and validated for an heterogeneous set of annotations in a large scale collection of patent documents.

Keywords: textual resources, TEI, annotations
Word count: 1978 (without references and figures)

1. Motivation

In the field of digital document representation, various proposals for the encoding of annotations in relation to an existing XML document have been proposed. When encoding such annotations, two different approaches can be considered: inline annotations, where the markup elements of the annotations are directly added into the annotated data, and standoff annotations, where the markup elements of the annotations reside in a location different from the location of the data being described.

Whereas the inline annotation has as main advantage its simplicity and easy maintenance, it also presents numerous drawbacks when implementing multiple types of annotations on the same source text, a common requirement in document engineering. In such cases, the standoff annotation mechanism has shown to be a better option (Dipper 2005).

The standoff annotation mechanism allows the definition of new layers of information on the top of the source textual data without disturbing existing ones, and are therefore appropriate for encoding multiple concurrent annotations on a source text.

The Text Encoding Initiative (TEI) is the major XML standard for encoding and documenting primary data (Przepiórkowski and Banski, 2009). However, most of the existing text encoding standards that allow the encoding of annotations, and in particular standoff annotations, have been proposed outside the TEI: the standards developed in the frame of ISO TC37/SC4 (Idé and Romary 2006), XCES (Idé, Bohomme and Romary 2000), TIGER-XML (Mengel and Lezius, 2000), PAULA (Dipper 2005) or XStandoff (Stührenberg and Jetka 2009).

There have been a few attempts to encode standoff annotations within the TEI (Boot 2009, Przepiórkowski 2009) (Banski and Przepiórkowski 2009) (Banski and Przepiórkowski 2010), or the TEIWiki page for "Stand-off

¹ http://wiki.tei-c.org/index.php/Stand-off_annotations

use cases". However, none of them offers a general framework for standoff annotations able to cope in a consistent way with a vast range of annotation types and to encompass the encoding standard proposed outside the TEI. These attempts only rely on specific combinations of already existing TEI elements for encoding the annotations in the context of specific projects.

The TEI guidelines disclose the basic referencing mechanisms for linking pieces of the source document with the annotations (see the TEI Guidelines: Chapter 16.9, 17.4 and 20.4), but they do not cover in a consistent manner the standoff annotations.

Our proposal aims at defining an extension of the TEI for representing standoff annotations fulfilling a set of comprehensive requirements (section 2). This supposes the introduction of additional elements described in section 3, allowing grouping and inter-operability with existing third party content annotation schema (section 4). The realization of the proposal is illustrated by heterogeneous automatic and manual annotations realized on a large scale collection of patent documents (section 5).

2. General encoding requirements

Our primary ambition is to provide an annotation framework going beyond the purpose of particular types of annotation, e.g. beyond linguistic annotation only, and beyond a particular family of documents, for instance texts only. We thus introduce the following requirements:

- **Expressiveness:** the framework must be able to encode heterogeneous types of standoff annotation models: multiple annotation layers, covering different content structures, references and segmentations.
- **Extensibility:** the framework is not limited to the encoding of standoff annotations of textual information, but support any type of source data, including audio, video, image, etc.
- **Uniformity:** the framework must provide a general mechanism that, even allowing different ways of encoding the annotation information, provides a

[Cancel](#)
[Next step](#)
 Recover the metadata from the PDF file



A Generic Formalism for Encoding Standoff Annotations in TEI

Javier Pose Rodriguez
European Patent Office
jposerodriguez@epo.org

Patrice Lopez
INRIA Saclay
patrice.lopez@inria.fr

Laurent Romary
INRIA Saclay
laurent.romary@inria.fr

Abstract

This article describes a proposal for a comprehensive and consistent encoding of standoff annotations in the frame of the TEI standard. The encoding requires the extension of the current TEI schema with three additional elements providing a generic and flexible structure for encoding standoff annotations in multiple layers or levels of annotations. The proposal, part of a bigger project still in progress, is implemented and validated for an heterogeneous set of annotations in a large scale collection of patent documents.

Keywords: textual resources, TEI, annotations
Word count: 1978 (without references and figures)

1. Motivation

In the field of digital document representation, various proposals for the encoding of annotations in relation to an existing XML document have been proposed. When encoding such annotations, two different approaches can be considered: inline annotations, where the markup elements of the annotations are directly added into the annotated data, and standoff annotations, where the markup elements of the annotations reside in a location different from the location of the data being described.

Whereas the inline annotation has as main advantage its simplicity and easy maintenance, it also presents numerous drawbacks when implementing multiple types of annotations on the same source text, a common requirement in document engineering. In such cases, the standoff annotation mechanism has shown to be a better option (Dipper 2005).

The standoff annotation mechanism allows the definition of new layers of information on the top of the source textual data without disturbing existing ones, and are therefore appropriate for encoding multiple concurrent annotations on a source text.

The Text Encoding Initiative (TEI) is the major XML standard for encoding and documenting primary data (Przeźiórkowski and Banski, 2009). However, most of the existing text encoding standards that allow the encoding of annotations, and in particular standoff annotations, have been proposed outside the TEI: the standards developed in the frame of ISO TC37/SC4 (Ide and Romary 2006), XCES (Ide, Bozhomme and Romary 2000), TIGER/XML (Mengel and Lezius, 2000), PAULA (Dipper 2005) or XStandoff (Stührenberg and Jetka 2009).

There have been a few attempts to encode standoff annotations within the TEI (Boot 2009, Przeźiórkowski 2009) (Banski and Przeźiórkowski 2009) (Banski and Przeźiórkowski 2010), or the TEIWiki page for "Stand-off

¹ http://wiki.tei-c.org/index.php/Stand-off_text_cases

use cases"¹. However, none of them offers a general framework for standoff annotations able to cope in a consistent way with a vast range of annotation types and to encompass the encoding standard proposed outside the TEI. These already exist annotations in TEI. These already exist annotations in TEI.

The TEI guidelines mechanisms fit with the annotation 16.9, 17.4 and 17.5 in a consistent manner the standard.

Our proposal aims at representing a comprehensive and consistent the introductory section 3, allow existing third party. The realization heterogeneous information on a large scale.

2. General Framework
Our primary framework goal of annotation, beyond a particular text only. We define the following characteristics:

- **Expressiveness:** heterogeneous multiple annotation structures, multiple levels of annotations.
- **Extensibility:** encoding information, including as a mechanism for encoding information.
- **Uniformity:** mechanism for encoding information.

A Generic Formalism for Encoding Standoff Annotations in TEI

Javier Pose Rodriguez

European Patent Office

jposerodriguez@epo.org

Patrice Lopez

INRIA Saclay

patrice.lopez@inria.fr

Laurent Romary

INRIA Saclay

laurent.romary@inria.fr

Abstract

This article describes a proposal for a comprehensive and consistent encoding of standoff annotations in the frame of the TEI standard. The encoding requires the extension of the current TEI schema with three additional elements providing a generic and flexible structure for encoding standoff annotations in multiple layers or levels of annotations. The proposal, part of a bigger project still in progress, is implemented and validated for an heterogeneous set of annotations in a large scale collection of patent documents.

Keywords: textual resources, TEI, annotations

Word count: 1978 (without references and figures)

1. Motivation

In the field of digital document representation, various proposals for the encoding of annotations in relation to an existing XML document have been proposed. When

use cases"¹. However, none of them offers a general framework for standoff annotations able to cope in a consistent way with a vast range of annotation types and to encompass the encoding standard proposed outside the

[Home](#)[Submit](#)[Browse](#)[Search](#)[Documentation](#)[My space](#)

[Submit](#) / [Conference papers](#)

The following metadata have been recovered from the PDF file :

Language, Title, Abstract, Keywords, Authors

[Home](#)[Submit](#)[Browse](#)[Search](#)[Documentation](#)[My space](#)**Title ***

Vous pouvez renseigner le titre en plusieurs langues : choisir la langue et cliquer sur + pour ajouter une nouvelle langue

A Generic Formalism for Encoding Standoff Annotations in TEI

English

**Abstract ***

Vous pouvez renseigner le résumé en plusieurs langues : choisir la langue et cliquer sur + pour ajouter une nouvelle langue

This article describes a proposal for a comprehensive and consistent encoding of standoff annotations in the frame of the TEI standard. The encoding requires the extension of the current TEI schema with three additional elements providing a generic and flexible structure for encoding standoff annotations in multiple layers or levels of annotations. The proposal, part of a bigger project still in progress, is implemented and validated for an heterogeneous set of annotations in a large scale collection of patent documents.

English

**Keywords**

Vous pouvez renseigner plusieurs mots-clés en plusieurs langues : choisir la langue et cliquer sur + pour ajouter un mot-clé dans la langue sélectionnée. Les caractères "," (virgule) et ";" (point-virgule) peuvent être utilisés pour séparer une liste de mots-clés.

textual resources

English



TEI

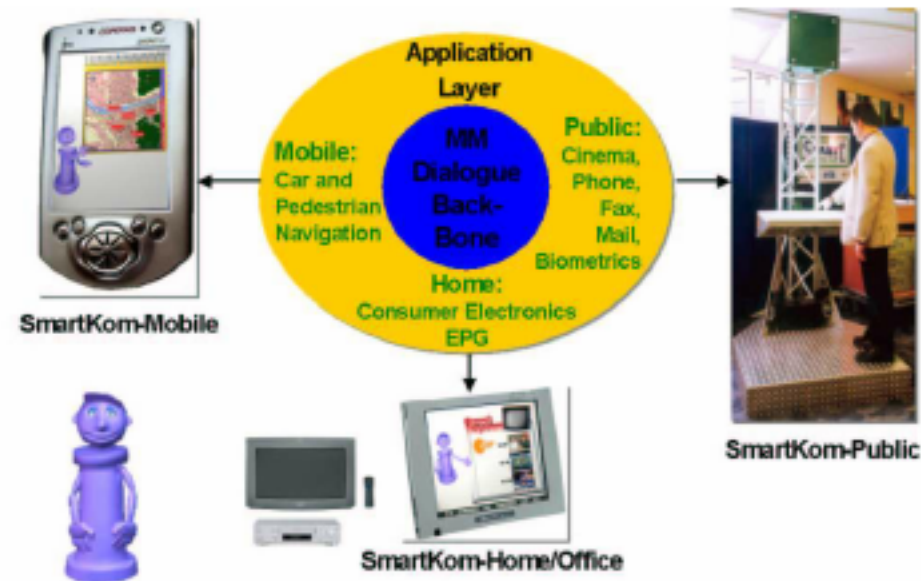
English



annotations

English





Figure

[Recommend](#)

[Comment](#)

Caption

Figure 1: SMARTKOM's dialogue backbone and applica-

0 Recommendations



Figure

[Recommend](#)

[Comment](#)

Caption

Figure 2: The force-feedback device developed in the MIAMM project. The display shows a view of a database using a timeline.

0 Recommendations

Automatically enrich PDF with interactive annotations

... 01300
1°40'N is
us struc-
: stability
e before
wever, at
s setting
near near
ace con-
with the
the pres-
in offset
n. It may
d zone.”
e role of
nown as
acteristic
ern trace,
w depres-
and the

[10] A general feature of propagating ridge segments along the northern MAR is the southward trend of all propagating segments between 26°30'N and the

Dannowski, A., I. Grevemeyer, C. R. Ranero, G. Ceuleneer, M. Maia, J. P. Morgan, and P. Gente (2010), Seismic structure of an oceanic core complex at the Mid-Atlantic Ridge, 22°19'N, *J. Geophys. Res.*, 115, B07106, doi:10.1029/2009JB006943.

linked to anagmatic cal... in a magmatically starved setting [\[Dannowski et al., 2010\]](#).

[11] The TAMMAR segment shows an hourglass shaped axial valley morphology that change from the segment center toward the segment ends. At 21°47'N the shallowest on-axis area of the segment is 2 km wide and ~3 km deep with walls ~600 m high. During a submersible study [\[Gente et al., 1996\]](#), two types of volcanism were identified on the inner floor: Isolated volcanoes and piled lava flows, flat lava lakes. They found that the isolated

Recognition of scientific entities

- Well known for biomedical entities, chemical formulas
- Experimenting with new kinds of entities:
 - ➔ physical measurements
 - ➔ astronomical entities

Identification and normalization of physical measures

COS-7 cells transfected with the indicated plasmids were lysed in Laemmli sample buffer or the lysis buffer mentioned above. E18.5 mouse brains (ICR) were homogenized in 20 mm HEPES (pH 7.4), 0.1 mm EDTA, 0.1 mm EGTA, 150 mm NaCl, 2 mm MgCl₂, 1 mm Na₃V₀4, 0.4 mm 4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride, 10 µg/ml leupeptin, and 1 mm dithiothreitol with a Teflon pestle homogenizer. The lysates or homogenates were centrifuged at 15,000 × g for 20 min, and the supernatants were used for immunoprecipitation of Cdk5 with anti-Cdk5 (C8) or anti-p35 (C19). In some cases, immunoprecipitation was performed with anti-Cdk5 (C8) or anti-p35 (C19) that had been cross-linked to protein A-Sepharose beads using the Pierce Crosslink IP kit according to the protocol of the manufacturer. The cell extracts were incubated with 1.5 µg of antibody and 20 µl of protein A-Sepharose beads and rotated overnight at 4 °C. The beads were washed with washing buffer (25 mm Tris-HCl (pH 7.5), 0.1 mm EDTA, 0.1 mm EGTA, 500 mm NaCl, 0.5% Nonidet P-40, and 1 mm dithiothreitol) five times. The kinase activity of Cdk5 was measured with histone H1 as a substrate in kinase buffer (10 mm MOPS (pH 6.8), 1 mm MgCl₂, 0.1 mm EDTA, and 0.1 mm EGTA) at 37 °C for 30 min. After SDS-PAGE, phosphorylation was visualized by autoradiography with an imaging plate.

Atomic value

quantity type: length

raw value: 1

raw unit name: mm

normalized value: 0.001

normalized unit name: m

quantified:

raw: dithiothreitol

normalized: dithiothreitol

<http://quantity.science-miner.com>

Identification and normalization of physical measures

Fifty-three journals were collected: 13 were eliminated from analysis, because they were incomplete, unclear or unreadable. 40 journals were analysed: 19 were journals of subjects of race Z (4 women and 15 men, 30 ± 10 years, 176 ± 7 cm, 70 ± 9 kg, 15 ± 5 % of fat mass, $\text{VO}_{2\text{max}}$: 50 ± 8 ml · kg⁻¹ · min⁻¹ and 21 of race A (6 women and 15 men, 40 ± 7 years, 176 ± 7 cm, 72 ± 10 kg, 18 ± 8 % fat mass, $\text{VO}_{2\text{max}}$: 58 ± 8 ml · kg⁻¹ · min⁻¹). Energy, macronutrients (CHO, fat and proteins) and liquid intakes were analysed.

Interval

quantity type: length

raw: from 169 to 183

raw unit name: cm

normalized: from 1.69 to 1.83

normalized unit: m

quantified:

raw: men

normalized: man

Identification and normalization of physical measures

ingestion may show adaptation to the natural seston concentrations (cf. KJØRBOE and SMITH, 1981).

The Oosterschelde estuary (SW Netherlands) has high particulate matter concentrations are between 270 mg.l^{-1} (WETSTEYN *et al.*, 1986). The seston is mainly composed of inorganic and refractory organic matter and phytoplankton forms only a minor fraction of the seston (SMAAL *et al.*, 1986). Earlier experimental studies (LINS and SMAAL, 1989) suggest that selectivities of mussels from the Oosterschelde are comparable to the efficiencies observed by WETSTEYN *et al.* (1981). In this paper, experimental results on the selective ingestion of algae by *Mytilus edulis* and *Cerastoderma edule* from the Oosterschelde estuary are reported. Experiments were carried out to establish the relation between the concentration of suspended particulate matter and the feeding efficiency. The impact of the selective ingestion on the food budget of the bivalves was

MATERIALS AND METHODS

Experiments with the blue mussel *Mytilus edulis* were carried out from January to March 1988. Experiments with the cockle *Cerastoderma edule* were carried out from March to April 1988. The animals were collected in the western part of the Oosterschelde, from sites near the low water tidal limit. They were transported to the field station of the Institute of Marine Biology, and stored in flowing sea water.

through a $125 \mu\text{m}$ sieve, and kept in stock suspensions of $10\text{-}20 \text{ g.l}^{-1}$. The carbon content of the silt was $2.7 \pm 1.7\%$, the C:N ratio was 19.6.

Selection experiments

The animals were kept in raceways, and fed with a mixture of the diatom *Phaeodactylum tricornutum* ($20 \cdot 10^3 \text{ cells.ml}^{-1}$) and silt. The algae and silt were continuously added to a flow of filtered natural sea water. The concentrations of suspended particulate matter (SPM) were set at a range from 5 to 90 mg.l^{-1} in the mussel experiments, and from 20 to 120 mg.l^{-1} in the cockle experiments. The experiments were carried out at *in situ* water temperatures, which varied between 5.1 and 7.9°C . The current velocity in the raceways was low ($<1 \text{ cm.s}^{-1}$) to prevent resuspension of faeces and pseudofaeces.

Before the start of the experiment animals were stored in filtered sea water for 24 hours. In each experiment 5 animals were fed with an experimental diet for a period of $20\text{-}24$ hours. At the end of that period faeces and pseudofaeces of each individual were collected separately. If necessary, faeces (and sand in the experiments with cockles) were sorted out of the pseudofaeces using the differences in density. Faeces and pseudofaeces were subsampled for various chemical analyses.

The suspended particulate matter in the diets, and pseudofaeces and faeces were analyzed for total dry weight, inorganic matter content, chlorophyll-*a* and phaeophytin-*a*. Dry weight was analyzed after drying at 70°C for 48 hours. Inorganic matter content was measured by determining ash content after ignition at 520°C for 4 hours. Chlorophyll-*a* and phaeophytin-*a* were determined by HPLC analysis.

Interval

quantity type: mass

raw: from 20 to 120

raw unit name: mg

normalized: from 0.00002 to 0.00012

normalized unit: kg

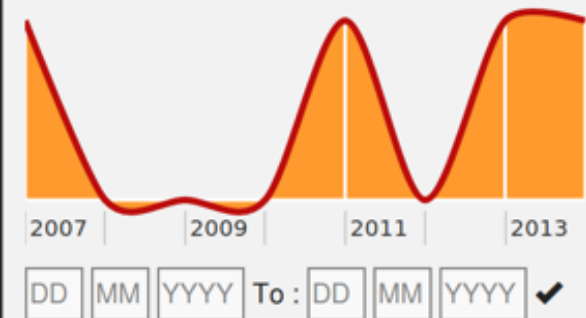
quantified (experimental):

raw: the cockle experiments

normalized: experiment

Searching quantities

publication_date + -



subject-headers + -



keywords + -

Swimming
Physical Sciences
Master athlete
Xterra

all fields all lang must search term Disambiguate + Quantities

length: 10-1000 m ✕

3 results - in 7 ms (server time)

Quantity search - form - free query - text ✕

length metre 10 1000 substance Disambiguate Parse +

PROTON shuttle MECHANISM in the TRANSITION STATE of LIPASE CATALYZED N-acylation of AMINO ALCOHOLS
Per-Olof Syren, Florian Le Joubioux, Yesmine Ben Henda, Thierry Maugard, Karl Hult, Marianne Graber - *ChemCatChem* - 2013
Abstract/Keywords ▾

Age-related changes in conventional ROAD versus OFF-ROAD TRIATHLON performance
Romuald Lepers, P.J. Stapley - *European Journal of Applied Physiology* - 2011
Abstract/Keywords ▾

PHYSICAL CHARACTERIZATION of the Potentially-Hazardous High-Albedo ASTEROID (33342) 1998 WT from Thermal-Infrared Observations
Alan W. Harris, Michael Mueller, Marco Delbó, Schelte J. Bus - *Icarus* - 2007
Abstract/Keywords ▾

Recognition of astronomical entities

050219	11:05:38.97	-40:41:02.6	1"9	ATCA 6A	2013 Feb. 10 12:24-10 14:27	0.82	1934-638	1104-445
				ATCA 6A	2013 Feb. 10 20:31-10 22:14	0.82	1934-638	1104-445

Notes. ^(a) The coordinates are the previous column. ^(c) The GRBs 020819B, 030528, 080319C, and 110918A (see below). ^(b) On-source integration time per snapshot within the time interval given in the previous column. ^(c) GRB 030528 has been remeasured on the original data of the first-epoch NTT data (Greiner et al. 2003), leading to a substantial error.

GRB hosts of our sample have a \sim extent smaller than about 1 arcsec; the exceptions are GRBs 020819B, 050219, 080319C, and 110918A (see below).

With ATCA, we observed our sample sources (project C2718; PI: J. Greiner) with the CFB 1M-0.5K mode in the 6 km configuration, providing 2048 channels per 2048 MHz continuum IF (1 MHz resolution) and 2048 channels per 1 MHz zoom band (0.5 kHz resolution). Most sources were observed over the full range of hour angles to ensure good uv -plane coverage.

Data analysis was performed using the standard software package MIRIAD (Sault et al. 1995), applying appropriate band-pass, phase, and flux calibrations. Substantial flagging had to be applied to remove radio frequency interference (RFI), removing up to 30% of the original data. Multifrequency synthesis images were constructed using robust weighting (robust = 0) and the full bandwidth between its flagged edges. The noise was determined by estimating the rms in emission-free parts of the cleaned map (using kvis).

2.2. VLA

We observed five sources at S band in B configuration (project 13B-017; PI: J. Greiner). The observations were performed in full polarization mode, with a total synthesized bandwidth of 2 GHz, centered at 3.0 GHz. We used standard amplitude and bandwidth calibration (observing 3C48, 3C147, or 3C295,

depending on the source), and a bright nearby phase calibrator for each of the targets (see Table 1). We reduced the data using the Common Astronomy Software Applications package (CASA; McMullin et al. 2007). The noise was determined as the rms in emission-free regions in the images.

The data reduction was problematic for four reasons: (i) the phase calibrators used were not optimal for S band in the observed configuration, with resolved structure and important closure errors; (ii) the strong radio frequency interference (RFI) was the main culprit for data flagging (see Table 2), except for GRB 080605 (see below); (iii) the presence of strong sources in the field that limited the dynamic range of the synthesized images (last column in Table 2); and (iv) significant gain variation due to variable power from geostationary satellites entering the analog signal path through the sidelobes of the antennas; this affects sources in the declination range from $-14^\circ 5' < \text{Dec} < +5^\circ 5'$ (e.g., Perley et al. 2015), thus necessitating $>60\%$ data flagging for GRB 080605.

3. Results

3.1. Radio flux measurements

We detect only one of our targets, the nearest one, namely GRB 020819B with a measured flux $F(3 \text{ GHz}) = 31 \pm 8 \mu\text{Jy}$. The peak of the radio emission is at RA (2000.0) = 23:27:19.50, Dec (2000) = +06:15:55.8 with an error of 0'3. This is 0'37

Semantic content enrichment

- Issues:
 - we don't have homogeneous data schema, taxonomy, etc.
 - we have very few classification information at article level
- Fully automatic content extraction/classification techniques:
 - Usage of Wikipedia/Wikidata
 - Usage of specialized scientific knowledge bases

Semantic content enrichment

<http://nerd.science-miner.com>

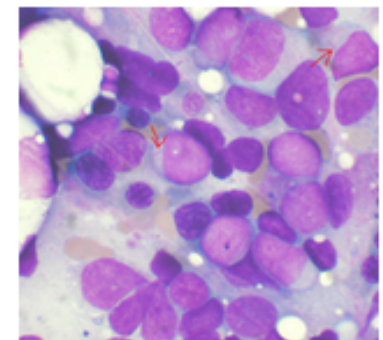
Development and maintenance of **LEUKEMIA** can be partially attributed to alterations in (anti)-apoptotic **GENE EXPRESSION**. **GENOME-WIDE TRANSCRIPTOME** analyses revealed that **89** apoptosis-associated **GENES** were differentially **EXPRESSED** between **PATIENT ACUTE MYELOID LEUKEMIA (AML) CD34(+) CELLS** and normal **BONE MARROW (NBM) CD34(+) CELLS**. Among these, **TRANSFORMING GROWTH FACTOR-B** activated **KINASE 1 (TAK1)** was strongly **UPREGULATED** in **AML CD34(+) CELLS**. **GENETIC** downmodulation or **PHARMACOLOGIC INHIBITION** of **TAK1** activity strongly impaired primary **AML CELL SURVIVAL** and cobblestone formation in **STROMAL** cocultures. **TAK1 INHIBITION** was mainly due to blockade of the **NUCLEAR FACTOR KB (NF-KB)** pathway, as **TAK1 INHIBITION** resulted in reduced levels of **P-IKBA** and **P65** activity. **OVEREXPRESSION** of a constitutive active variant of **NF-KB** partially rescued **TAK1**-depleted **CELLS** from **APOPTOSIS**. Importantly, **NBM CD34(+) CELLS** were less sensitive to **TAK1 INHIBITION** compared with **AML CD34(+) CELLS**. **KNOCKDOWN** of **TAK1** also **SEVERELY** impaired **LEUKEMIA** development in vivo and **PROLONGED** overall survival in a **HUMANIZED XENOGRAFT MOUSE MODEL**. In conclusion, our results indicate that **TAK1** is frequently **OVEREXPRESSED** in **AML CD34(+) CELLS**, and that **TAK1 INHIBITION** efficiently **TARGETS LEUKEMIC** stem/progenitor **CELLS** in an **NF-KB**-dependent manner.

ACUTE MYELOID LEUKEMIA

Normalized: Acute myeloid leukemia

Domains: Medicine

conf: 0.958136066797794



Acute myeloid leukemia (AML) is a cancer of the myeloid line of blood cells, characterized by the rapid growth of abnormal white blood cells that accumulate in the bone marrow and interfere with the production of normal blood cells. AML is the most common acute leukemia affecting adults, and its incidence increases with age. Although AML is a relatively rare disease, accounting for roughly 1.2% of cancer deaths in the United States, its incidence is expected to increase as the population ages.

Reference: 

Semantic content enrichment

With the advent of **WHITE MAN** in **NORTH AMERICA** and his consequent modification of the environment by **LUMBERING** and clearing for **FARMING**, **COYOTES** have been extending their range. As they have extended their range, on the fringes of their **NEWLY ACQUIRED TERRITORIES**, **ANIMALS** which are difficult to identify have frequently been captured. In the **SOUTH**, as often as not, these are called **RED WOLVES**, in the **NORTHEAST**, coydogs. In both **PARTS OF THE COUNTRY** these **ANIMALS** occur where **COYOTES** have moved into areas that formerly were **INHABITED** by small races of **WOLF**. **COINCIDENT** also with these shifts in distribution has been an upward revision in the reported weights for **COYOTES**. **ELIMINATING** a few out-sized individuals, give a range of **18-30** pounds for typical western **COYOTES**, while gives a range of **23-50** pounds for **MICHIGAN COYOTES**. The latter **OVERLAPS** with weights of a **LONG SERIES** of **WOLVES** from **ALGONQUIN PROVINCIAL PARK** (unpublished data from the **ONTARIO** Department of Lands and **FORESTS**) and, as a result, size alone becomes a less useful criterion in **DISTINGUISHING** between **WOLVES** and coyotes.

In the following discussion, since **CANIS LUPUS**, the **WOLF**, and **CANIS** latrans, the **COYOTE**, are both composite **SPECIES**, these names as used in the text **REFER** to each **SPECIES** as a unit. When a particular **SUBSPECIES** is referred to, a trinomial is used, as **CANIS LUPUS LYCAON**. **CANIS NIGER**, the **RED WOLF**, is usually considered to include three **SUBSPECIES**. Their status is uncertain, and **CANIS NIGER** as used in the present work **REFERS** to the typical form, **C. n. NIGER**, and to those southeastern **POPULATIONS**, presently called **C. n. gregoryi**, which show no evidence of **HYBRIDIZATION** and which were collected from well outside the range of latrans. **CANIS FAMILIARIS**, the **DOG**, presents no problem because, in spite of its variability, it is monotypic.

The present study was **UNDERTAKEN** because attempts to identify **SKULLS** of the northeastern **POPULATION** of rather **LARGE-SIZED** members of the **GENUS** **CANIS** bogged down in a mass of **OVERLAPPING** characters. It was then decided that before such fringe **POPULATIONS** could be identified we needed to know what, if any, combinations of characters reliably separated known **CANIS LUPUS**, latrans, and **FAMILIARIS**, particularly if size were eliminated as a character. This part of the work will be described in detail in **SECTION I**.

Wikidata properties
part of the KB

CANIS LUPUS

Normalized: **Gray wolf**

Domains: **Animals**

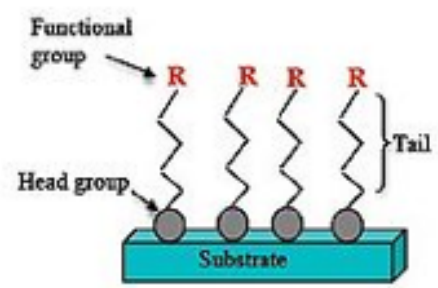


conf: 0.8799286110348177



The **gray wolf** or **grey wolf** (*Canis lupus*), also known as the **timber wolf** or **western wolf**, is a **canid** native to the wilderness and remote areas of **Eurasia** and **North America**. It is the largest **extant** member of its family, with males averaging , and females . Like the **red wolf**, it is distinguished from other **Canis** species by its larger size and less pointed features, particularly on the ears and muzzle. Its winter fur is long and bushy, and predominantly a mottled gray in color, although nearly pure white, red, or brown to black also occur. , 37 **subspecies** of *C. lupus* are recognised by **MSW3**.

genus	Canis
tribus	Canini
classis	Mammalia
regnum	Animalia

Demo: <http://keyterm.science-miner.com>

term	score	entities
self-assembled monolayers	0.1649	<p>self-assembled monolayers</p> <p>conf: 0.2191</p> <p>Self-assembled monolayer:</p> <p>A self assembled monolayer (SAM) is an organized layer of amphiphilic molecules in which one end of the molecule, the "head group" shows a specific, reversible affinity for a substrate. SAMs also consist of a tail with a functional group at the terminal end as seen in Figure 1.</p>  <p>The diagram illustrates a self-assembled monolayer (SAM) on a substrate. It shows four molecules, each with a grey circular head group attached to a blue rectangular substrate. From each head group, a zigzag line representing a tail extends upwards. At the end of each tail is a red 'R' labeled as a functional group. A bracket on the right side of the tails is labeled 'Tail'. A 'W' icon is in the top right corner of the diagram area.</p>
dispersion forces	0.0845	<p>dispersion forces</p> <p>conf: 0.9889</p> <p>Van der Waals force:</p> <p>In physical chemistry, the van der Waals force (or van der Waals interaction), named after Dutch scientist Johannes Diderik van der Waals, is the sum of the attractive or repulsive forces between molecules (or between parts of the same molecule) other than those due to covalent bonds or to the electrostatic interaction of ions with one another or with neutral molecules. The term includes:</p>  <p>A photograph of a black gecko clinging to a light-colored, textured wall. The gecko is positioned vertically, facing right, with its tail pointing downwards. This image is used to illustrate the concept of van der Waals forces, which allow the gecko to adhere to the wall. A 'W' icon is in the top right corner of the image area.</p>
density functional	0.0209	<p>density functional</p> <p>conf: 0.9889</p>  <p>A partial view of a molecular model or simulation, showing a dark, textured surface with several bright blue or cyan spots. A 'W' icon is in the top right corner of the image area.</p>

Semantic content enrichment

Demo: <http://traces1.saclay.inria.fr/inria>

The screenshot displays the HAL search interface. At the top left is the Inria logo with the tagline "INVENTEURS DU MONDE NUMÉRIQUE" and the HAL logo with "archives-ouvertes.fr". On the top right, there are navigation options: "simple - complex - NL".

The search bar contains "search term" and shows "76,133 results - in 257 ms (server time)". A dropdown menu for "publication_date" is open, showing a histogram of results. Below it are date selection fields: "DD MM YYYY to DD MM YYYY". Another dropdown menu for "subject-headers" is open, showing a circular sunburst chart of subject categories. At the bottom left, a "keywords" dropdown shows "physical sciences (872)" and "france (556)".

The main content area displays a search result for "ird-00968855 - Intercomparison of four remote-sensing-based ENERGY BALANCE methods to retrieve SURFACE EVAPOTRANSPIRATION and WATER STRESS of IRRIGATED FIELDS in SEMI-ARID CLIMATE" by J. Chirouze et al. - 31.12.2014. The abstract is displayed with semantic enrichment: "Abstract: Instantaneous EVAPOTRANSPIRATION RATES and surface WATER STRESS levels can be deduced from REMOTELY SENSED SURFACE TEMPERATURE DATA through the SURFACE ENERGY budget. Two FAMILIES of methods can be defined: the contextual methods, where STRESS levels are scaled on a given IMAGE between hot/dry and cool/wet PIXELS for a particular VEGETATION cover, and single-pixel methods, which evaluate LATENT HEAT as the residual of the surface ENERGY BALANCE for one PIXEL independently from the others. Four models, two contextual (S-SEBI and a modified TRIANGLE method, named VIT) and two single-pixel (TSEB, SEBS) are applied over one GROWING SEASON (December-May) for a 4 KM x4 km IRRIGATED AGRICULTURAL AREA in the SEMI-ARID northern Mexico. Their performance, both at local and SPATIAL standpoints, are compared relatively to ENERGY BALANCE DATA acquired at seven locations within the area, as well as an UNCALIBRATED soil-vegetation-atmosphere transfer (SVAT) MODEL forced with local in situ DATA including observed IRRIGATION and RAINFALL amounts. STRESS levels are not always well retrieved by most models, but S-SEBI as well as TSEB, although slightly biased, show good performance. The drop in MODEL PERFORMANCE is observed for all MODELS when VEGETATION is".

On the right side, there is a "SURFACE ENERGY" section with a blue header. It includes "Domains: Engineering" and "conf: 0.88". Below this is an image of a water droplet on a surface. The text below the image reads: "'Surface energy'" quantifies the disruption of intermolecular bonds that occurs when a surface is created. In the [[physics]] of [[solid]]s, surfaces must be intrinsically less energetically favorable than the bulk of a material, otherwise there would be a driving force for surfaces to be created, removing the bulk of the material (see [[sublimation (chemistry)|sublimation]]). The surface energy may therefore be defined as the

Semantic content enrichment

Amedeo Napoli (200)
Rémi Gribonval (180)
Jérôme Euzenat (177)
Olivier Festor (174)
François Chaumette (168)
Cordelia Schmid (163)
Laurent Romary (162)
Olivier OD Devillers (158)

concepts

Algorithm (781)
Scientific modelling (569)
Data (338)
Computer simulation (322)
Equation (303)
Computer software (290)
Mathematical optimization (259)
Design (251)
Conceptual model (243)
Homogeneity and heterogeneity (235)

keyterms

Machine learning (34,169)
Educational software (32,548)
Mathematical optimization (1,103)
Learning (938)
Data mining (787)
Par (score) (772)
Machine (733)
Wireless sensor network (687)
Simulation (640)
Robot (617)

hal-01242157 Journal articles

Abstract: **NEGATIVE FEEDBACK** circuits are a **RECURRENT MOTIF** in regulatory **BIOLOGICAL** networks, strongly linked to the **EMERGENCE** of oscillatory **BEHAVIOR**. The theoretical analysis of the existence of oscillations is a difficult problem and typically involves many constraints on the monotonicity of the activity functions. Here, we study the occurrence of periodic solutions in an **N-DIMENSIONAL** class of negative **FEEDBACK** systems defined by **SMOOTH VECTOR FIELDS** with a window of not necessarily monotonic activity. Our method consists in **CIRCUMSCRIBING** the **SMOOTH** system by two **PIECEWISE LINEAR** ones, each admitting a **PERIODIC SOLUTION**. It can then be shown that the **SMOOTH NEGATIVE FEEDBACK** system also has a **PERIODIC ORBIT**, inscribed in the **TOPOLOGICAL SOLID TORUS** constructed from the two piecewise linear orbits. The interest of our approach lies in first, adopting a general class of **FUNCTIONS**, with an nonmonotonicity window, which permits a better fitting between theoretical models and experimental **DATA**, and second, establishing a more accurate location for the **PERIODIC SOLUTION**, which is useful for computational purposes in **HIGH DIMENSIONS**. As an **ILLUSTRATION**, a model for the "repressilator" synthetic system is analyzed and compared to **REAL DATA**, and shown to admit a **PERIODIC ORBIT**, for a range of activity **FUNCTIONS**.

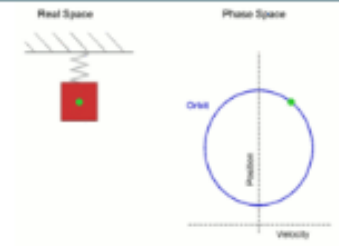
Keywords: Piecewise linear systems, negative feedback circuits, periodic oscillations, Poincaré maps AMS subject classifications 34, 92

PERIODIC ORBIT

Normalized: **Orbit (dynamics)**

Domains: **Physics**

conf: 0.37



In **mathematics**, in the study of **dynamical systems**, an **orbit** is a collection of points related by the **evolution function** of the dynamical system. The orbit is a subset of the **phase space** and the set of all orbits is a **partition** of the phase space, that is different orbits do not intersect in the phase space. Understanding the properties of orbits by using topological method is one of the objectives of the modern theory of dynamical systems.

Reference: **W**



Lagrange–Schwarz Waveform Relaxation domain decomposition methods for linear and nonlinear quantum wave problems

Xavier Antoine, Emmanuel Lorin - Applied Mathematics Letters - 2016

Abstract/Keywords

Semantic content enrichment

Amedeo Napoli (200)
Rémi Gribonval (180)
Jérôme Euzenat (177)
Olivier Festor (174)
François Chaumette (168)
Cordelia Schmid (163)
Laurent Romary (162)
Olivier OD Devillers (158)

concepts

Algorithm (781)
Scientific modelling (569)
Data (338)
Computer simulation (322)
Equation (303)
Computer software (290)
Mathematical optimization (259)
Design (251)
Conceptual model (243)
Homogeneity and heterogeneity (235)

keyterms

Machine learning (34,169)
Educational software (32,548)
Mathematical optimization (1,103)
Learning (938)
Data mining (787)
Par (score) (772)
Machine (733)
Wireless sensor network (687)
Simulation (640)
Robot (617)

hal-01242157 Journal articles

Abstract: **NEGATIVE FEEDBACK** circuits are a **RECURRENT MOTIF** in regulatory **BIOLOGICAL** networks, strongly linked to the **EMERGENCE** of oscillatory **BEHAVIOR**. The theoretical analysis of the existence of oscillations is a difficult problem and typically involves many constraints on the monotonicity of the activity functions. Here, we study the occurrence of periodic solutions in an **N-DIMENSIONAL** class of negative **FEEDBACK** systems defined by **SMOOTH VECTOR FIELDS** with a window of not necessarily monotonic activity. Our method consists in **CIRCUMSCRIBING** the **SMOOTH** system by two **PIECEWISE LINEAR** ones, each admitting a **PERIODIC SOLUTION**. It can then be shown that the **SMOOTH NEGATIVE FEEDBACK** system also has a **PERIODIC ORBIT**, inscribed in the **TOPOLOGICAL SOLID TORUS** constructed from the two piecewise linear orbits. The interest of our approach lies in first, adopting a general class of **FUNCTIONS**, with a non-monotonicity window, which permits a better fitting between theoretical models and experimental **DATA**, and second, establishing a more accurate location for the **PERIODIC SOLUTION**, which is useful for computational purposes in **HIGH DIMENSIONS**. As an **ILLUSTRATION**, a model for the "repressilator" synthetic system is analyzed and compared to **REAL DATA**, and shown to admit a **PERIODIC ORBIT**, for a range of activity **FUNCTIONS**.

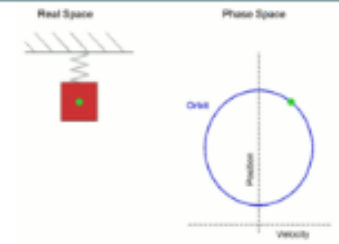
Keywords: Piecewise linear systems, negative feedback circuits, periodic oscillations, Poincaré maps AMS subject classifications 34, 92

PERIODIC ORBIT

Normalized: **Orbit (dynamics)**

Domains: **Physics**

conf: 0.37



In **mathematics**, in the study of **dynamical systems**, an **orbit** is a collection of points related by the **evolution function** of the dynamical system. The orbit is a subset of the **phase space** and the set of all orbits is a **partition** of the phase space, that is different orbits do not intersect in the phase space. Understanding the properties of orbits by using topological methods is one of the objectives of the modern theory of dynamical systems.

Reference: **W**



Lagrange–Schwarz Waveform Relaxation domain decomposition methods for linear and nonlinear quantum wave problems




Xavier Antoine, Emmanuel Lorin - Applied Mathematics Letters - 2016

Abstract/Keywords ▼

Query disambiguation

all fields ▾ all lang ▾ must ▾ concrete pump sensor ⚙️ Disambiguate +

3,185 results - in 303 ms (server time)

<input type="checkbox"/>	concrete	Concrete is a composite construction material, composed of cement (commonly Portland cement) and other cementitious materials such as fly ash and slag cement , aggregate (generally a coarse aggregate made of gravel or crushed rocks such as limestone , or granite , plus a fine aggregate such as sand), water and chemical admixtures.		W
<input type="checkbox"/>	concrete pump	A concrete pump is a tool used for transferring liquid concrete by pumping . There are two types of concrete pumps. The first type of concrete pump is attached to a truck . It is known as a trailer-mounted boom concrete pump because it uses a remote-controlled articulating robotic arm (called a boom) to place concrete with pinpoint accuracy . Boom pumps are used on most of the larger construction projects as they are capable of pumping at very high volumes and because of the labour saving nature of the placing boom. They are a revolutionary alternative to truck-mounted concrete pumps.		W
<input type="checkbox"/>	pump	A pump is a device used to move fluids, such as liquids , gases or slurries . A pump displaces a volume by physical or mechanical action. Pumps fall into three major groups: direct lift , displacement , and gravity pumps. Their names describe the method for moving a fluid.		W

Machine learning has also a performance impact on...

- Disambiguation and deduplication
 - ➔ author, affiliation, article record matching
- Recommendation systems
 - ➔ but require usage data

TDM: opportunity of new activities for information scientists

- Imagine new workflows and applications integrating these techniques
- Curation of research data
- Curation of training data for data structuring
 - Training data is the new oil !
- Testing and evaluating the tools
 - independent tests, benchmarking
 - end-to-end tests
 - A/B tests and testing methodologies

Tools & Demos

- GROBID: <https://github.com/kermitt2/grobid>
† demo: <http://grobid.science-miner.com>
- GROBID-Quantity: <https://github.com/kermitt2/grobid-quantities>
† demo: <http://quantity.science-miner.com>
- (N)ERD: <https://github.com/kermitt2/grobid-ner> (partial!)
demo: <http://nerd.science-miner.com>
- Keyterm extraction: not yet on GitHub
demo: <http://keyterm.science-miner.com>
- anHALytics: <https://github.com/anHALytics>
demo: <http://traces1.saclay.inria.fr/Inria>



Apache 2.0

