



**Chantier d'usage ISTEEX**

**Inria-Alpage**

**(équipe ALMAnaCH au 01/01/2017)**

# Équipe




Achraf Azhar

ISTEX 



Patrice Lopez

 science-miner



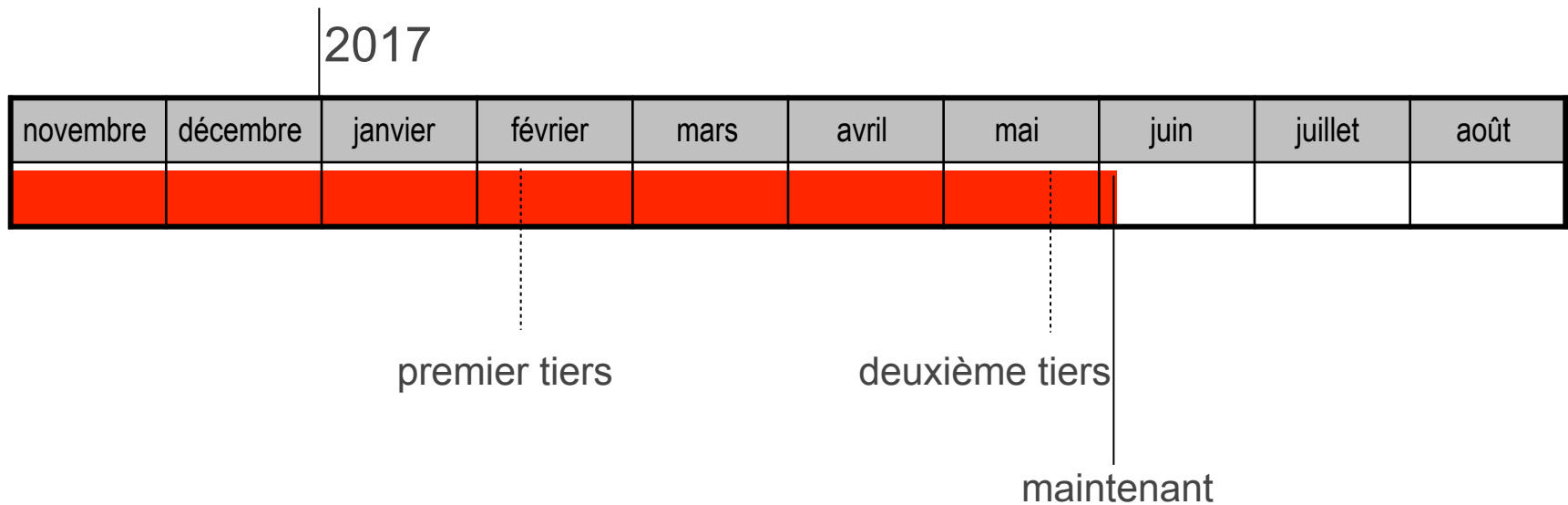
Luca Foppiano



Laurent Romary

# Avant propos

- Nos travaux pour le chantier d'usage :



# Le TDM scientifique aujourd'hui

- Fouille de textes scientifiques : rendre exploitable l'information contenue dans la littérature scientifique et technique
  - ➔ Amélioration des outils de recherche d'information
  - ➔ Construction de bases de connaissances
  - ➔ Production automatique d'hypothèses scientifiques

# Le TDM scientifique aujourd'hui

- Fouille de textes scientifiques : rendre exploitable l'information contenue dans la littérature scientifique et technique
  - ➔ Amélioration des outils de recherche d'information
  - ➔ Construction de bases de connaissances
  - ➔ Production automatique d'hypothèses scientifiques
- Des résultats bien concrets :
  - ➔ Outils de recherche documentaire
  - ➔ Services bibliographiques/documentaires
  - ➔ Découverte et développement de médicaments (facteur 4)

# Le TDM scientifique aujourd'hui

- Fouille de textes scientifiques : rendre exploitable l'information contenue dans la littérature scientifique et technique
  - ➔ Amélioration des outils de recherche d'information
  - ➔ Construction de bases de connaissances
  - ➔ Production automatique d'hypothèses scientifiques
- Des résultats bien concrets :
  - ➔ Outils de recherche documentaire
  - ➔ Services bibliographiques/documentaires
  - ➔ Découverte et développement de médicaments (facteur 4)
- Mais très couteux :
  - ➔ Expertises pour la construction de règles/patterns, bases de connaissances structurées exploitables à grandes échelles

# Nos travaux en TDM



- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)

# Nos travaux en TDM



Apache 2.0

- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)
- Extraction de termes, concepts et catégories clefs : KeyTerm



# Nos travaux en TDM



Apache 2.0

- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)
- Extraction de termes, concepts et catégories clefs : KeyTerm
- Désambiguïsation d'entités connues : (N)ERD (Wikipedia/data)

# Nos travaux en TDM



Apache 2.0

- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)
- Extraction de termes, concepts et catégories clefs : KeyTerm
- Désambiguïsation d'entités connues : (N)ERD (Wikipedia/data)
- Ingestion documentaire TEI - Pub2TEI



Apache 2.0

# Nos travaux en TDM

- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)
- Extraction de termes, concepts et catégories clefs : KeyTerm
- Désambiguïsation d'entités connues : (N)ERD (Wikipedia/data)
- Ingestion documentaire TEI - Pub2TEI
- Exploitation de HAL comme corpus pour le TDM et l'observation de l'activité scientifique française (anHALytics)

# Nos travaux en TDM



- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)
- Extraction de termes, concepts et catégories clefs : KeyTerm
- Désambiguïsation d'entités connues : (N)ERD (Wikipedia/data)
- Ingestion documentaire TEI - Pub2TEI
- Exploitation de HAL comme corpus pour le TDM et l'observation de l'activité scientifique française (anHALytics)
- Extractions spécialisées/nomenclaturées : extraction d'entités biotech, astronomiques, chimiques, de mesures scientifiques

# Nos travaux en TDM



- Structuration automatique de documents brutes (PDF) : GROBID (ResearchGate, CERN, NASA, HAL, Mendeley, etc.)
- Extraction de termes, concepts et catégories clefs : KeyTerm
- Désambiguïsation d'entités connues : (N)ERD (Wikipedia/data)
- Ingestion documentaire TEI - Pub2TEI
- Exploitation de HAL comme corpus pour le TDM et l'observation de l'activité scientifique française (anHALytics)
- Extractions spécialisées/nomenclaturées : extraction d'entités biotech, astronomiques, chimiques, de mesures scientifiques
- Déduplication (*matching*) d'entités scientifiques

# Objectifs du chantier d'usage ISTEEX

- Utilité du corpus ISTEEX pour la conception, l'industrialisation et la montée en charge d'annotateurs
  - ➔ annotations sur un grand échantillon - correspondant à 10 à 20% de l'ensemble du corpus ISTEEX
  - ➔ croiser les disciplines,
  - ➔ utiliser des sources réelles (PDF éditeurs)
- Encodage TEI stand-off des annotations
- Démonstrateur exploitant ces annotations basé sur notre infrastructure anHALytics

# Priorisation

- L'équipe ISTEEX R&D a déjà entrepris depuis 2014 le ré-entraînement et l'intégration de GROBID dans ISTEEX
  - ➔ 1 million de PDF traités en 24h (9 threads) - 11,5 PDF/s
- L'équipe ISTEEX R&D expérimente depuis début 2016 les modules d'extraction de termes, concepts et catégories clefs (grobid-keyterm) et de reconnaissance et extraction d'entités (N)ERD
- Par complémentarité, pour ce chantier d'usage, nous nous concentrons sur l'annotateurs grobid-quantities et son expérimentation sur le corpus ISTEEX

# Identification et normalisation des mesures physiques

- Origines :
  - ➔ besoin identifié pour les brevets
  - ➔ demande du JPL (NASA), utilisateur de GROBID
- Besoin transversal à toutes les disciplines scientifiques
- Solutions existantes limitées, propriétaires, par règles manuelles



démo

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

# Identification et normalisation des mesures physiques

- Extension grobid-quantities
- Notre approche :
  - ➔ machine learning, cascades LC-CRF
  - ➔ reconnaissance valeurs, intervalles, listes
  - ➔ couverture d'un maximum de types de mesures
  - ➔ normalisation en unités SI (JSR-363, ISO 80000, ...)
  - ➔ annotation de texte, XML mais aussi PDF
  - ➔ open source Apache 2
- 1000 mots par seconde (1 thread)
- Reconnaissance des substances quantifiées

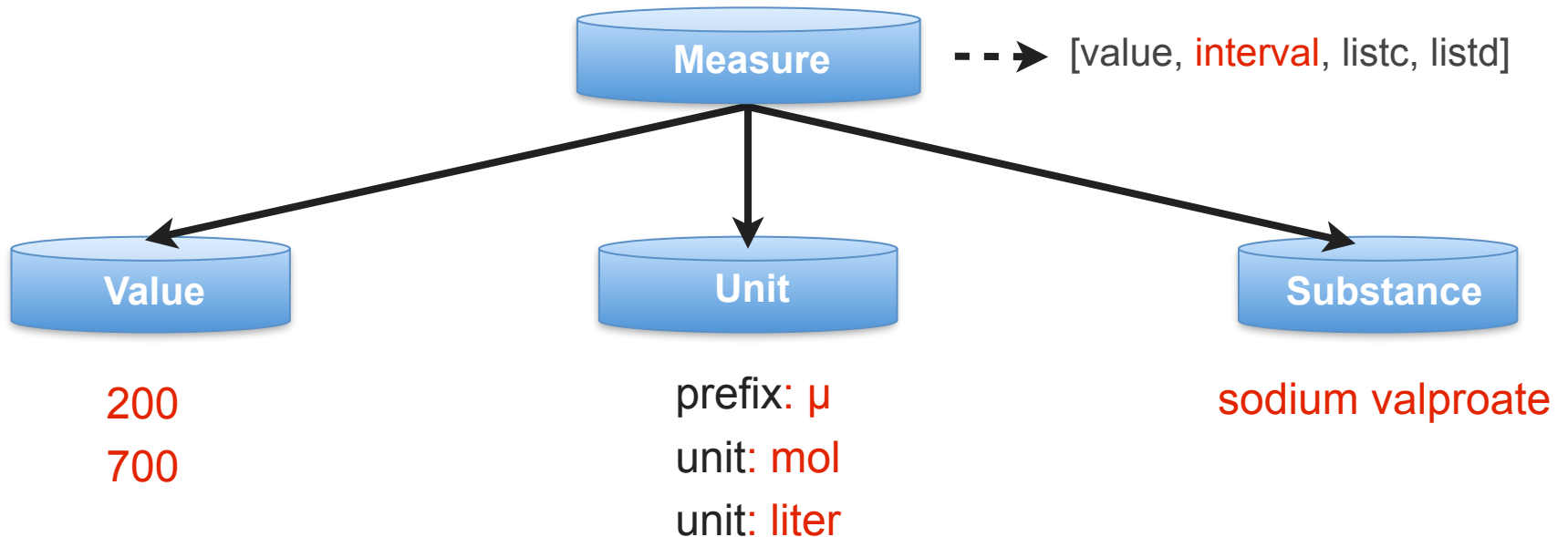
démo

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

JPL NASA integration <https://github.com/khundman/marve>

# Modèles CRF dans grobid-quantities

... and 200–700  $\mu\text{mol/l}$  of sodium valproate with ...



type: amount of substance

normalization: [0.2, 0.7]

normalized unit (SI): mol/m<sup>3</sup>

normalized: sodium valproate ( 650844)

from doi:10.1186/1475-2832-3-13

### Subjects and Study Design

All subjects gave full informed consent, and both studies were approved by the ethics committee at the University of Alberta. Healthy controls were examined using a detailed, but non-standardized, psychiatric interview. They were excluded if there was any personal history, or immediate family history, of psychiatric disorder. For patients, diagnoses were made using DSM-IV criteria for Bipolar Disorder Type I or Type II following detailed psychiatric interview, with additional information being available in almost all cases from long-term psychiatric clinic records. They also had to be taking a dose of either lithium or valproate which maintained their blood levels within the ranges of 0.4-1.2 mmol/l for lithium and 200-700  $\mu\text{mol/l}$  for sodium valproate. Serum lithium and valproate levels were also measured on the day of MRS scanning. Other medications taken by the patient were noted. In the second part of the study the same criteria were used, except that only patients meeting diagnostic criteria for Bipolar Disorder Type I were included, and they had to be on sodium valproate monotherapy. This was done to examine Bipolar Type I patients in more detail, and to remove a possible confounding variable. All patients had to be euthymic for the previous 2 months, as determined

Insert: [see definition of euthymia](#)

by structured clinical interviews and validated questionnaires

### Interval

quantity type: concentration

raw: from 200 to 700

raw unit name:  $\mu\text{mol/l}$

normalized: from 0.2 to 0.7

normalized unit:  $\text{mol/m}^3$

quantified (experimental):

raw: sodium valproate

normal

# Recherche documentaire par quantités

- Lucene : recherche de valeurs discrètes dans des intervalles
- Uwe Schindler & *BKD tree structures* dans Lucene :
  - ➔ possibilité de recherche spatiale pour objets 2D, voir 3D
  - ➔ fonctionne aussi en 1D : recherche d'intervalles par intervalles et *scoring* spécifique

# Recherche documentaire par quantités



# Recherche documentaire par quantités

- Lucene : recherche de valeurs discrètes dans des intervalles
- Uwe Schindler & *BKD tree structures* dans Lucene :
  - ➔ possibilité de recherche spatiale pour objets 2D, voir 3D
  - ➔ fonctionne aussi en 1D : recherche d'intervalles par intervalles et *scoring* spécifique



# Recherche documentaire par quantités

- Lucene : recherche de valeurs discrètes dans des intervalles
- Uwe Schindler & *BKD tree structures* dans Lucene :
  - ➔ possibilité de recherche spatiale pour objets 2D, voir 3D
  - ➔ fonctionne aussi en 1D : recherche d'intervalles par intervalles et *scoring* spécifique
  - ➔ des évolutions très récentes : Lucene 04/2016, Solr 08/2016, Elasticsearch 01/2017

# Recherche documentaire par quantités

- Lucene : recherche de valeurs discrètes dans des intervalles
- Uwe Schindler & *BKD tree structures* dans Lucene :
  - ➔ possibilité de recherche spatiale pour objets 2D, voir 3D
  - ➔ fonctionne aussi en 1D : recherche d'intervalles par intervalles et *scoring* spécifique
  - ➔ des évolutions très récentes : Lucene 04/2016, Solr 08/2016, Elasticsearch 01/2017
- Nouvelles perspectives en recherche d'information scientifiques et en text mining
- Exemple : recherche de documents par critères de positionnement 3D d'entités astronomiques dans l'univers

# Objectifs du chantier d'usage ISTEEX

- Utilité du corpus ISTEEX pour la conception, l'industrialisation et la montée en charge d'annotateurs
  - ➔ annotations sur un grand échantillon - correspondant à 10 à 20% de l'ensemble du corpus ISTEEX
  - ➔ croiser les disciplines,
  - ➔ utiliser des sources réelles (PDF éditeurs)
- Encodage TEI stand-off des annotations
- Démonstrateur exploitant ces annotations basé sur notre infrastructure anHALytics

# SOURCES

arXiv.org

PMC

HAL  
archives-ouvertes.fr

OAI-PMH

PDF

TFI

XMI

ISTEX

...

Crossref

\$

DEDUPLICATION

## KNOWLEDGE BASE

Entités de la recherche

GRAPHDB

## TEI

## INDEX

## PERSISTANCE

PDF

OCR

XML/TEI

ASSETS

ANNEXES

## ANNOTATEURS

OCR

GROBID

(N)ERD

KEYTERM

GROBID-QU

BIOGROBID

...

# Ingestion de ressources ISTEEX

- Utilisation des catégories WoS et ScienceMetrix pour la sélections de domaines riches en mentions de quantités
- Test et travaux actuels sur un échantillon
- Difficultés :
  - ➔ limitations de l'approche par ZIP : volume, efficacité, stabilité, évolution
  - ➔ suggestions : passer à OAI-PMH avec la définition d'ensembles et sous-ensembles pertinents
  - ➔ suggestions : séparer l'harvesting des métadonnées du téléchargement des PDF

démo

questions?

questions?



questions?

questions?

questions?

questions?

questions?

# Le TDM : des opportunités de nouvelles activités pour les documentalistes et bibliothécaires

- Imaginer des nouveaux workflows et de nouvelles applications intégrant ces techniques
- Curation de données de la recherche
- Curation de données d'apprentissage pour les outils d'annotations
- Tester et évaluer ces outils
  - ➔ test indépendant, benchmarking
  - ➔ test end-to-end
  - ➔ test A/B et méthodologies

# Les blocages

ISTEX

- Droit d'accès à la littérature scientifique et technique
- Légalité de la fouille de texte (US/UK/JP vs FR/DE)
- Besoin de couverture, données à jour (+ 1,5M articles/an)

GROBID

- Difficulté d'exploitation du format PDF / pauvreté et incohérence des metadonnées
- *“the information that I can extract from an article, at least for me, is not quite the information I want”*, Shreejoy Tripathy (neuroscientifique)

# Les défis scientifiques et techniques

- Gestion de la volumétrie : million de documents, milliards d'annotations, graphe avec milliard de noeuds
- Gestion de l'incertitude et du bruit: les meilleures techniques de fouille de textes font des erreurs, beaucoup d'erreurs...
- *Machine learning* : domaines et scénarios plus complexes et ouverts que le web marchand et les “add clicks”
- Prise en compte des méthodologies, opinions, réseaux, etc. propres à chaque discipline



# Tirer profit de la fouille de textes dans ISTEK

- Un très grand volume de données afin de neutraliser le bruits, les erreurs d'extractions, etc.
- Décloisonner les disciplines: croiser des champs/ disciplines
- Travail de normalisation, nettoyage des données extraites
- Capturer le langage spécialisé

# Approche

- Les approches par apprentissage automatique dominant toutes les compétitions en extraction d'information de textes scientifiques

## How to improve the performance of a ML application?

- Get better algorithm? +
- Get better features? +++
- Get better data? ++++++

(Leon Bottou)

# Capter les concepts connus : (N)ERD

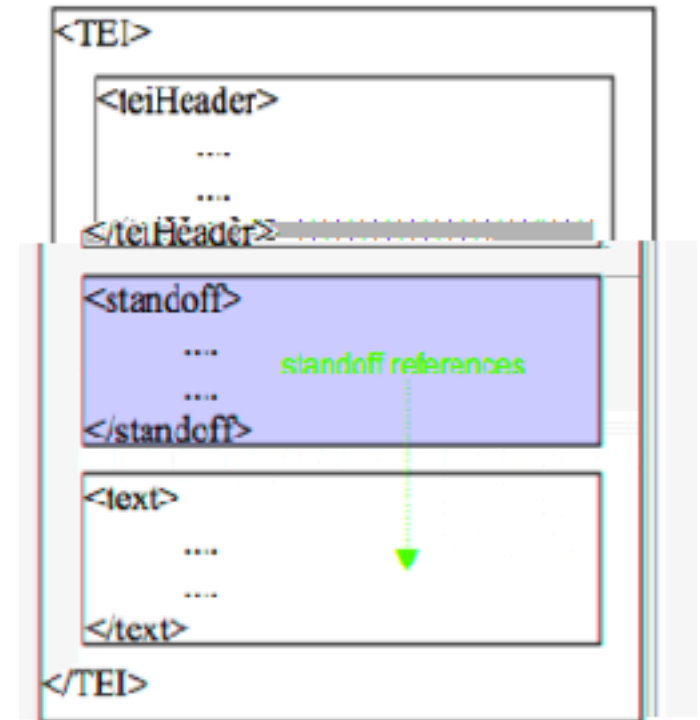
- Identification et résolution des entités avec des bases de connaissances : Wikipedia, FreeBase (Wikidata)
  - Couverture immédiate : > 4M de “concepts”, ~150M formes lexicales (pour l’anglais)
  - Pas d’autres contraintes d’expertise humaine en ingénierie des connaissances
  - Pas de contrainte de domaine
  - Multilingue
  - Catégorisation suivant les catégories Wikipedia/type Freebase
- ➔ mais ne couvre que des entités connues!

# Reconnaissance d'entités scientifiques liées à des nomenclatures

- Nomenclature: un système de nommage permettant de produire des termes et des concepts dans un champ particulier des sciences
- Terminologie et nomenclature = réduire l'impact de la langue naturelle
- Les entités ne peuvent toutes être énumérées par avance car les nomenclatures sont génératives.
  - ➔ PubChem du NIH par exemple reporte plus de 100 millions de formules chimiques dans le corpus brevets mondial
  - ➔ Central en chimie, biologie, astronomie, etc.
  - ➔ Pas de recherche professionnelle d'infos dans ces domaines sans prendre en compte ces termes

# Codage TEI d'annotations automatiques

- Persistance et réutilisation des annotations
- Nécessité de schémas détaillés d'annotations, agnostique -> TEI
- Annotations *standoff* en TEI avec offset
- Gestion de multiples annotations concurrentes, cumulables
- Positions préservées



# Travail du chantier d'usage ISTEEX

1. Annotations spécialisées en information scientifique et technique sur un grand échantillon - correspondant à 10 à 20% de l'ensemble du corpus ISTEEX
  - (N)ERD : entités connus / catégorisation
  - BIO-GROBID/BEAST : biotech
  - CHEMICAL-GROBID : chimie
  - grobid-quantities : mesures physiques
2. Encodage TEI stand-off des annotations
3. **Démonstrateur exploitant ces annotations basé sur notre infrastructure anHALytics**

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

The screenshot displays the HAL website interface. At the top left is the HAL logo with the URL <http://traces1.saclay.inria.fr>. The top right corner features the text "simple - complex - NL". Below the header is a search bar with the text "search term" and a "Disamb./Expand" button. A search result is shown for the paper "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. Chhouze et al., dated 01/12/2014. The paper's abstract is visible, discussing the use of remotely sensed surface temperature data and the surface energy budget to estimate evapotranspiration and water stress levels. The interface includes various filters on the left, such as "publication\_date" (with a date range selector), "subject-headers", and "keywords". A circular visualization of subject headers is also present. On the right, there is a section titled "SURFACE ENERGY" with a domain of "Engineering" and a confidence score of 0.89, accompanied by an image of a water droplet on a surface.

simple - complex - NL

+ add new facet

publication\_date

DD MM YYYY to DD MM YYYY ✓

subject-headers

keywords

- physical sciences (872)
- france (666)

76,133 results - In 257 ms (server time)

Search: search term Disamb./Expand

ird-00988855 - Intercomparison of four remote-sensing-based ENERGY BALANCE methods to retrieve SURFACE EVAPOTRANSPIRATION and WATER STRESS of IRRIGATED FIELDS in SEMI-ARID CLIMATE  
J. Chhouze et al. - 01/12/2014

ird-00988855

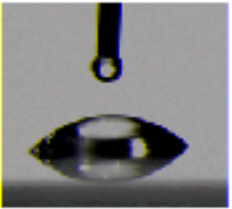
JOURNAL ARTICLE

Jonas Chhouze  
G. Boulet  
L. Jorin  
R. Flaizel  
J. C. Rodriguez  
J. Ezzafer  
S. Er Raki  
G. Bigeard  
O. Martin  
J. Gerstlitz-Payan  
C. Watts  
C. Chelbourn

**Abstract:** Intercomparison of SURFACE EVAPOTRANSPIRATION RATE 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 KMx4 km IRRIGATED AGRICULTURAL AREA in the SEMI-ARID northern Mexico. Their performance, both at local and SEVERAL 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

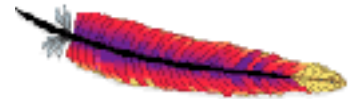
**SURFACE ENERGY**

Domain: Engineering  
conf: 0.89



"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 can therefore be defined as the

# Liens



**Apache 2.0**

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