Zeitschrift: Horizons : le magazine suisse de la recherche scientifique
Herausgeber: Fonds National Suisse de la Recherche Scientifique
Band: 27 (2015)
Heft: 105

Artikel: 200 ans de littérature en 0,4 seconde
Autor: Bischofberger, Mirko
DOI: https://doi.org/10.5169/seals-771908

200 ans de littérature en 0,4 seconde

La lecture lente a vécu: des ordinateurs avalent des millions de livres en un rien de temps. Et nous proposent leur propre analyse. *Par Mirko Bischofberger*

Tout a commencé avec des données. Beaucoup trop nombreuses et trop complexes. En 1940, le jésuite Roberto Busa s'attelle à produire un index complet des 11 millions de mots retrouvés dans les écrits théologiques de Thomas d'Aquin. Une entreprise gigantesque, pour laquelle une vie entière ne suffirait pas. Mais le père Busa a une idée: se faire aider par une machine. Il trouve le soutien de Thomas Watson, le fondateur d'IBM, et après trois décennies vient à bout de son projet: 56 volumes et 70 000 pages. L'Index Thomisticus est la première œuvre à permettre une recherche simple et rapide dans les contenus d'un corpus entier.

La montagne et le sport

La numérisation s'immisce aujourd'hui dans toutes les sciences humaines. «La linguistique et la littérature sont les principales intéressées, explique Martin Volk, professeur de linguistique computationnelle à l'Université de Zurich. Un accès numérique à leurs données de recherche leur permet d'étayer ou d'invalider certaines hypothèses à l'aide de chiffres et de statistiques.» Dans le cadre de son propre projet de recherche Text+Berg, le spécialiste a numérisé les 250 volumes du Club Alpin Suisse (CAS) parus depuis 1864. «Ce matériel est une mine d'informations. Il montre comment la façon d'envisager les montagnes a évolué avec le temps. Décrites autrefois comme des objets d'exploration, elles sont vues aujourd'hui comme un terrain d'entraînement sportif. Le terme ‹compétition› est, par exemple, beaucoup plus fréquent que par le passé.»

A l'Université de Genève, des chercheurs veulent numériser une partie de la Bibliotheca Bodmeriana, une collection exceptionnelle de 150 000 œuvres littéraires recouvrant trois millénaires en 80 langues. On y trouve, entre autres, le plus ancien manuscrit de l'Evangile de Jean datant du IIe siècle ainsi que les originaux des contes des frères Grimm.

Le citoyen à l'aide

Mais numériser des ouvrages est une tâche pénible. «Il faut couper les livres à la main avant de scanner chaque page séparément, détaille Martin Volk, qui a numérisé dans son projet plus de 120 000 pages. On a alors uniquement des images, mais pas de texte.» La reconnaissance des textes est effectuée par des programmes qui identifient les lettres dans les images et les transforment en mots. «Mais le taux d'erreurs reste encore assez élevé, notamment avec des écrits anciens du XIXe siècle.». Dans son projet, ce taux était de douze erreurs par page. Il a tout fallu vérifier à la main.

«L'outil Ngram joue un rôle pionnier pour les sciences humaines.»

Martin Volk a alors développé un système de correction en ligne qui a permis à des volontaires d'éliminer les erreurs, sous forme de jeu. Ce projet de science citoyenne a séduit les membres du CAS. «Grâce à leur aide, nous avons pu effectuer plus de 250 000 corrections en six mois.» Le corpus numérique est aujourd'hui presque correct à 100%. Une fois numérisés, les textes peuvent être facilement archivés et consultés, une chose «impossible pour des documents anciens, rares ou difficilement accessibles», souligne le chercheur.

Freud, Einstein, Darwin

Google Books est sans doute l'archive de ce genre la plus célèbre et la plus complète. Sa recherche en texte intégral permet de parcourir les stocks des bibliothèques universitaires de Harvard, Stanford et New York. Certaines bibliothèques européennes sont également intégrées, comme celle de l'Université d'Oxford ou la Bayerische Staatsbibliothek.

Ce gigantesque corpus a donné naissance en 2010 à Google Ngram, une application web qui analyse la fréquence d'un mot ou d'une suite de mots dans tous les

◂ Pages 15 et 16. Les deux nous parlent de ce que nous mangeons: un regard jeté dans notre frigo et une visualisation analytique des arômes. La couleur indique les catégories, la taille des nœuds représente leur fréquence dans les recettes, et les liens le nombre de composants aromatiques qu'ils partagent.
Images : Valérie Chételat (p. 15); Yong-Yeol Ahn (p. 16)

ouvrages dès 1800 scannés par Google. Elle permet d'étudier des événements historiques, comme l'abolition de l'esclavage, mais aussi d'observer l'évolution linguistique de certains mots au sein d'une langue, ou encore la popularité des personnalités: les célébrités scientifiques que sont Freud, Einstein ou Darwin apparaissent toutes très fréquemment dans la littérature, mais Freud est cité deux fois plus souvent depuis 1950.

«Ngram n'est qu'un exemple de ce qu'on peut faire aujourd'hui avec les données culturelles numérisées», remarque son créateur Jean-Baptiste Michel, chercheur à l'Université Harvard. Les sciences humaines numériques sont aujourd'hui inconcevables sans Ngram, dont Martin Volk confirme le «rôle pionnier».

La culture du SMS

La numérisation de la littérature n'est qu'une approche d'analyse linguistique. «Avec nos ordinateurs et téléphones, nous saisissons plus de textes numériques que jamais», souligne Elisabeth Stark, du Séminaire de langues romanes de l'Université de Zurich. Rien qu'en 2013, plus de 100 millions de SMS ont été envoyés chaque jour en Allemagne. «Ces textes ne sont presque jamais imprimés, mais font partie de notre culture linguistique», poursuit la chercheuse. Avec son projet du Fonds national suisse Sms4science, elle étudie les caractéristiques linguistiques de la communication par SMS en Suisse.

La langue raccourcie des SMS obéit aux même lois que celles du langage parlé.

Pour accéder à ces données, Elisabeth Stark et ses collègues ont invité les utilisateurs de portables en Suisse à envoyer une copie de leurs SMS à un numéro gratuit. «Nous avons ainsi pu collecter environ 26 000 messages.» Son équipe s'intéresse notamment aux ellipses linguistiques, autrement dit aux omissions de mots, comme dans les expressions «Arrive bientôt» ou «T'appelle». Afin de découvrir pourquoi le sujet est omis dans ces exemples, les chercheurs ont analysé tous les SMS en français et en allemand. Résultat: ces omissions sont beaucoup plus rares qu'on ne l'imaginait et obéissent aux mêmes lois que le langage parlé quotidien. «Cela contredit l'impression que l'on a souvent lorsqu'on regarde des SMS isolés, explique Elisabeth Stark. D'où la nécessité de disposer d'une grande quantité de données.»

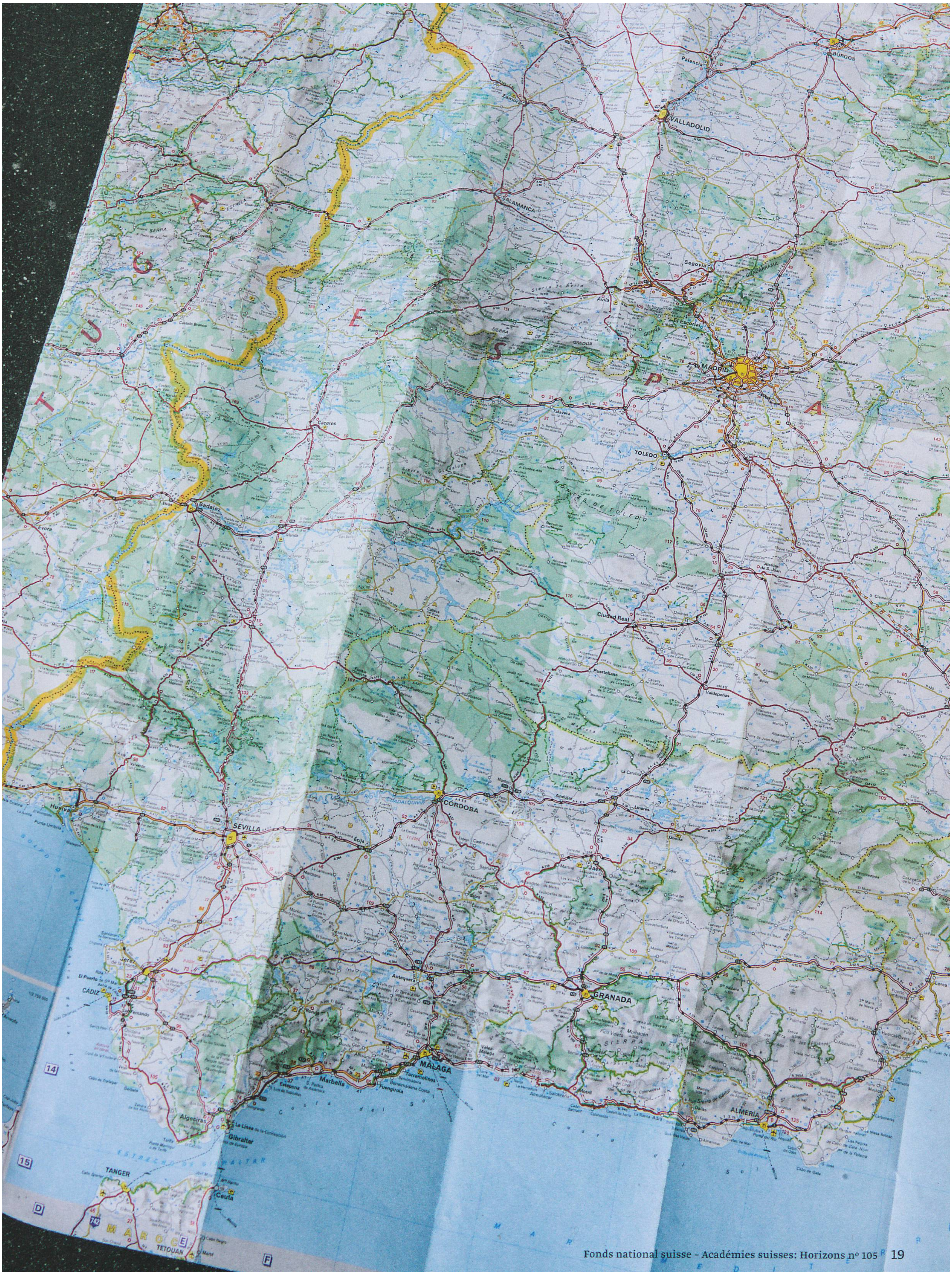
Accéder aux données

Les sciences humaines numériques permettent d'analyser la littérature et la langue à l'aide de chiffres, qui ont toujours été la marque des sciences exactes. Ils décrivent les schémas quantitatifs et les relations avec une précision dont les mots sont rarement capables. La prochaine génération de chercheurs en sciences humaines travaillera sur des données, comme le font les bioinformaticiens depuis la fin du XXe siècle. «Ce domaine profitera de l'augmentation massive de la quantité de textes numérisés, opine Martin Volk. Comme le séquençage du génome, qui a conduit à la bioinformatique, la numérisation de notre langue et de notre littérature fera inévitablement bientôt partie intégrante des sciences humaines.»

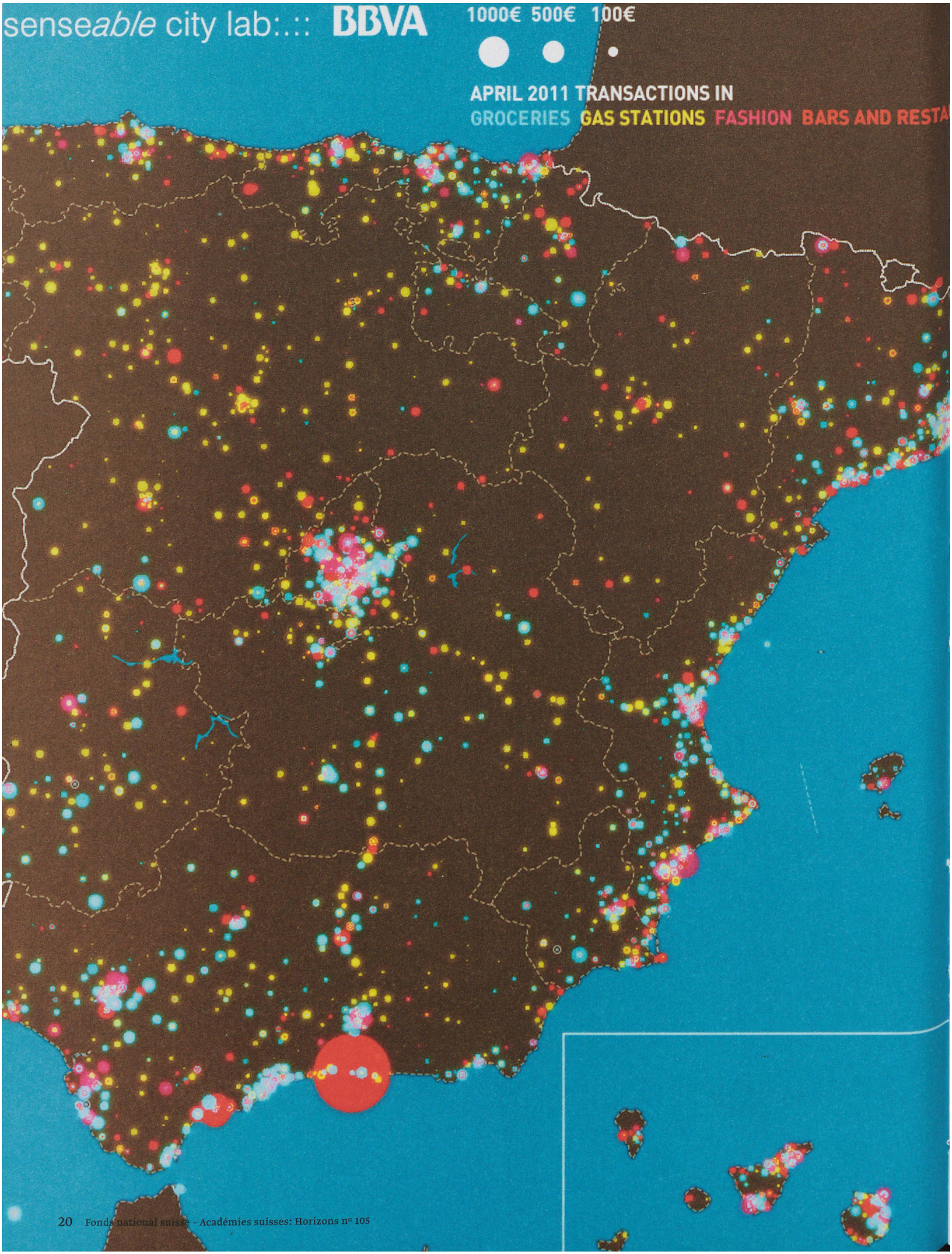
Des chercheurs comme Martin Volk et Elisabeth Stark ne représentent que le début d'une nouvelle ère dans la recherche. «Malheureusement, les ressources allouées aux sciences humaines numériques sont limitées pour l'instant, en Suisse», déplore Martin Volk. «Dans toute l'Université de Zurich, par exemple, il n'y a pas encore de chaire d'humanités numériques alors qu'il serait grand temps», renchérit Elisabeth Stark. Mais les deux chercheurs semblent trouver encore plus important de pouvoir accéder aux principaux réservoirs de données. «Des initiatives européennes d'envergure existent, mais la Suisse n'en fait malheureusement pas partie pour l'instant», regrette Elisabeth Stark. Jean-Baptiste Michel abonde dans son sens: «Pouvoir accéder aux données constitue le moteur essentiel!»

Mirko Bischofberger est collaborateur scientifique du FNS.

J.-B. Michel et al., Quantitative Analysis of Culture Using Millions of Digitized Books. Science, 2011



BURGOS
Palencia
VALLADOLID
SALAMANCA
Segovia
Ávila
MADRID
TOLEDO
Cáceres
Badajoz
Évora
Ciudad Real
Valdepeñas
CORDOBA
SEVILLA
Huelva
Écija
Jaén
GRANADA
Antequera
MALAGA
Torremolinos
Fuengirola
Marbella
Estepona
Algeciras
Gibraltar
CÁDIZ
TANGER
Ceuta
TETOUAN
ALMERIA
14
15
D
F


senseable city lab:.:: BBVA
1000€ 500€ 100€
APRIL 2011 TRANSACTIONS IN
GROCERIES GAS STATIONS FASHION BARS AND RESTA