Découverte d'indicateurs de classement dans le Web des données

Cyril De Runz, Arnaud Giacometti, Béatrice Markhoff, Arnaud Soulet

Université de Tours, LIFAT, Blois
firstname.lastname@univ-tours.fr

Résumé. Analyser l'impact d'entités au sein de leur domaine est fondamental pour le comprendre. A cette fin, il est essentiel de disposer d'indicateurs numériques fins retranscrivant les spécificités du domaine. Cet article propose une approche pour découvrir automatiquement des indicateurs d'impact pour classer les entités. Bien que l'approche soit transdisciplinaire, les indicateurs de classement identifiés doivent néanmoins disposer d'une sémantique intradisciplinaire. Pour cela, notre approche s'appuie sur les bases de connaissances du Web des données, pas seulement pour faciliter le calcul opérationnel des indicateurs mais aussi pour profiter de leur transparence et de la formalisation explicite de leur sémantique. L'hypothèse simple mais centrale de ce travail est que chaque répartition inégalitaire d'une quantité engendre un indicateur de classement pertinent. A cette fin, nous utilisons le coefficient de Gini pour identifier dans Wikidata les propriétés produisant des indicateurs de classement significatifs.

1 Introduction

Quel est le classement des peintres les plus importants ? Il n'est pas difficile de trouver une réponse à cette question, mais trouver la bonne s'avère plus compliqué. Par exemple, la table 1 présente les 10 peintres les plus importants selon trois sites. Ces classements s'accordent sur une partie des peintres (en gras dans les tableaux), mais certains artistes jugés comme primordiaux dans un seul classement sont occultés par les deux autres (en italique). Cela n'a évidemment rien de surprenant puisque ces classements découlent de l'avis subjectif d'une personne [1] ou d'un ensemble de personnes [2]. Le troisième classement consiste à classer les peintres par popularité en mesurant le nombre de lectures de leur article dans une encyclopédie en ligne [3] (voir l'indice en troisième colonne du classement `artcyclopedia.com` signifiant, par exemple, que la page consacrée à Salvador Dalí a reçu environ moitié moins de visites que celle de Pablo Picasso). L'utilisation d'un indicateur numérique est rassurant mais soulève aussi des questions de pertinence qui seront justement abordées dans cet article.

Notre question initiale concernant le classement des peintres pourrait être étendue à d'autres domaines : quels sont les processus biologiques les plus prépondérants ? Quelles sont les villes

1. `www.theartwolf.com/articles/most-important-painters.htm`
2. `www.ranker.com/list/best-painters-of-all-time/ranker-art`
3. `www.artcyclopedia.com/mostpopular.html`

theartwolf.com

Rang	Peintre
1	**Pablo Picasso**
2	*Giotto Di Bondone*
3	**Leonardo da Vinci**
4	*Paul Cézanne*
5	**Rembrandt**
6	*Diego Velasquez*
7	*Wassily Kandinsky*
8	**Claude Monet**
9	Caravaggio
10	*Joseph M. W. Turner*

ranker.com

Rang	Peintre
1	**Leonardo da Vinci**
2	Vincent van Gogh
3	Michelangelo
4	**Rembrandt**
5	**Pablo Picasso**
6	**Claude Monet**
7	Caravaggio
8	*Johannes Vermeer*
9	*Raphael*
10	Salvador Dalí

artcyclopedia.com

Rang	Peintre	Indice
1	**Pablo Picasso**	100
2	Vincent van Gogh	77
3	**Leonardo da Vinci**	65
4	**Claude Monet**	56
5	Salvador Dalí	47
6	*Henri Matisse*	43
7	**Rembrandt**	41
8	*Andy Warhol*	34
9	*Georgia O'Keeffe*	33
10	Michelangelo	32

TAB. 1: Trois classements (parmi tant d'autres) des peintres les plus importants.

prédominantes sur le plan économique ? etc. En effet, il est régulier d'avoir à comparer l'importance d'entités entre elles (par exemple, des villes) sur des critères ciblés (pour les villes, la comparaison peut porter sur l'économie ou la culture) notamment pour conduire des politiques publiques (Behn, 2003) ou scientifiques (Weinberg, 2000). Dans ce cas, même s'il faut rester prudent (Muller, 2018), le recours à un indicateur numérique est un atout car il établit de facto un ordre strict partiel rendant aisé la comparaison de deux entités. De plus, la définition de l'indicateur donne un sens au classement (par exemple, classement par popularité pour un nombre de vues). Le choix de l'indicateur est discutable, mais pas son classement sous-jacent si les données exploitées sont correctes et complètes, et bien sûr, accessibles. Enfin, la méthode de classement est reproductible par d'autres personnes et sur d'autres données ce qui est notamment utile pour mesurer des évolutions temporelles. Dans ce contexte, notre objectif ambitieux vise à définir une méthode de découverte d'indicateurs de classement qui soit *transdisciplinaire* dans le sens où une approche concerne simultanément plusieurs domaines. Malheureusement, les sous-objectifs opposés de la transdisciplinarité des approches et de leur impact intradisciplinaire constitue un défi majeur. Pour faire sens, les indicateurs numériques sont construits manuellement à l'intention d'un type d'entités et uniquement pour une discipline visée. Par exemple, même si le nombre de citations en bibliométrie est transposable en webométrie par le nombre d'hyperliens entrants, ces deux indicateurs sont distincts. Identifier automatiquement de telles quantités pertinentes quel que soit le domaine (allant par exemple, de la peinture à l'urbanisme en passant par les processus biologiques) est une tâche non-triviale.

Cet article propose une méthode de découverte automatique d'indicateurs de classement en s'appuyant sur le Web des données. Premièrement, par sa simplicité opérationnelle, sa transparence et sa richesse sémantique, nous plaidons dans la section 2 pour l'usage du Web des données comme ressource fondamentale pour la définition d'indicateurs. Nous pointons aussi la confusion sémantique qui limite l'usage des indicateurs utilisés en recherche d'information. Deuxièmement, nous définissons dans la section 3 un indicateur de classement comme un comptage d'une quantité inégalement répartie entre les entités à comparer. Nous montrons alors comment détecter une telle inégalité à l'aide du coefficient de Gini en calculant le seuil minimum à satisfaire. A notre connaissance, aucun travail de la littérature n'avait établi ce lien entre inégalité et indicateur de classement. Enfin, nous mettons à l'épreuve dans la section 4 notre méthodologie sur Wikidata qui est une base de connaissances libre éditée de manière collaborative. Par sa transdisciplinarité, les indicateurs de classement découverts concernent tous les domaines mais en restant cohérents par rapport à leur domaine visé.

2 Web des données comme jumeau numérique

Les données numérisées sont la condition sine qua non à la construction généralisée d'indicateurs numériques. Par exemple, les premières méthodes bibliométriques ont été proposées dès le début du 20ème siècle pour étudier l'impact d'un domaine ou d'un groupe de chercheurs. Pourtant, il a fallu attendre les années 1970 pour un essor de la bibliométrie grâce à l'informatisation des bases de données notamment au sein de l'Institute for Scientific Information (ISI) (Garfield, 1972). L'importance des données est tout aussi centrale en infométrie où chaque nouvelle source de données est l'occasion de produire de nouveaux indicateurs. Typiquement, le classement des peintres de `artcyclopedia.com` s'appuie sur la fréquence de visite des pages associées à chacun des peintres. Cette méthode relève de l'analyse des données d'usage du Web qui est une technique classique en webométrie (Thelwall et al., 2005). Sur un plan bibliométrique, il serait aussi possible de comptabiliser le nombre de publications consacrées à chaque artiste pour mesurer l'influence des peintres. Bien sûr, toutes ces données sont pertinentes pour mesurer l'influence d'un peintre d'un point de vue informationnel, mais cela n'est pas directement lié à leur impact disciplinaire. A l'image des nombres de publications ou de citations pour un scientifique, il serait préférable de s'appuyer sur le nombre d'oeuvres peintes ou de peintres influencés. Plus généralement, un indicateur numérique devrait mesurer des quantités ayant une signification intradisciplinaire tangible dans le monde réel.

Dans cette perspective, les bases de connaissances du Web des données (Berners-Lee et al., 2001) (aussi dénommé par « données ouvertes liées » ou « Web sémantique ») forment une large source de données prometteuse. En effet, le Web des données concerne tous les domaines de la connaissance et il favorise la publication de données structurées ce qui facilite leur exploitation pour des traitements statistiques. La facette « données liées » incite à l'interconnexion des données en référençant les entités du monde réel avec des identifiants uniques (appelés URI pour *Uniform Resource Identifier*). Cela évite des étapes fastidieuses de traitement des données comme la désambiguïsation. Par ailleurs, le Web des données s'inscrit aussi dans un mouvement d'ouverture des données. Cette ouverture améliore naturellement la couverture des indicateurs et donc leur fiabilité, mais elle augmente aussi la transparence qui fait défaut aux indicateurs construits à partir de bases de données privées. Enfin, la facette « Web sémantique », sûrement la plus importante, souligne la richesse des données en formalisant ce que les données représentent et ce qu'on peut en faire. Alors qu'un hyperlien relie uniformément deux pages du Web, les propriétés qualifient le sens de l'interconnexion entre deux entités. Elles sont porteuses de la spécificité disciplinaire désirée. Mieux, les ontologies structurent et standardisent ces propriétés de sorte qu'il est possible de raisonner sur ces bases de connaissances. A l'inverse, il faut tenir compte des faiblesses des bases de connaissances du Web des données. Si les informations renseignées sont globalement fiables, elles souffrent souvent d'incomplétude (Zaveri et al., 2016; Razniewski et al., 2016) qui entrainent des biais de représentativité (Soulet et al., 2018). Dans notre cas, cela risque de fausser les indicateurs numériques.

Comme le Web des données est avant tout une partie du Web, une première tentation serait de lui appliquer les méthodes webométriques. De manière simple, une partie des travaux peut être transposée en considérant une entité comme une page Web identifiée par une URI et une propriété comme un hyperlien du sujet vers l'objet. Il est aussi possible de se restreindre à une partie des propriétés correspondant à une ontologie particulière comme Friend of a Friend (Stuart, 2014). Une seconde tentation est de s'appuyer sur les mesures définies dans le cadre de la recherche d'information qui peut être vue comme un domaine connexe à l'infométrie. En

effet, il y a un parallèle naturel entre les méthodes bibliométriques analysant les citations et les méthodes de recherche d'information analysant les liens (Kleinberg, 1999; Page et al., 1999) – qui ont inspiré des scores similaires dans le Web sémantique (Ding et al., 2004). Dans ces travaux, le score d'importance d'une entité (article ou page) augmente avec son degré entrant de pointeurs (citations ou liens). Le principe populaire en recherche d'information de pondérer l'apport d'un pointeur suivant l'importance de sa source a même été envisagé auparavant par certains indicateurs bibliométriques comme celui proposé par Pinski et Narin (1976). Concernant les travaux du Web sémantique, les scores issus de la webométrie ou de la recherche d'information considèrent donc uniformément toutes les entités sans bénéficier des spécificités propres à leur domaine. Par exemple, ces scores adisciplinaires permettent de comparer l'importance d'un peintre avec celle d'une ville. Cela s'explique par la sémantique des liens entre entités, c'est-à-dire des propriétés, qui est complètement ignorée. En outre, ces scores en agrégeant des informations hétérogènes (à travers les propriétés variées) n'ont pas vraiment de signification dans le monde réel. Pire, deux choix de représentation des connaissances différents conduiront à des scores différents.

3 Indicateur de classement pertinent = forte inégalité

Avant de poursuivre, précisons quelques notations. Une base de connaissances sur un ensemble de propriétés $\mathcal{P}$ et un ensemble d'entités $\mathcal{E}$ est un ensemble de faits $\mathcal{K} \subseteq \mathcal{E}\times\mathcal{P}\times\mathcal{E}$. Nous écrivons les faits sous la forme $\langle s, p, o\rangle \in \mathcal{K}$, où s est le sujet, p est la propriété et o est l'objet. Par exemple, $\langle \texttt{Guernica}, \texttt{creator}, \texttt{Picasso}\rangle$ signifie que le tableau Guernica a été peint par Pablo Picasso [4]. Etant donnée une propriété p, $\langle s, p^{-1}, o\rangle \in \mathcal{K}$ signifie que $\langle o, p, s\rangle \in \mathcal{K}$ où p^{-1} est la propriété inverse de p (e.g., on a $\langle \texttt{Picasso}, \texttt{creator}^{-1}, \texttt{Guernica}\rangle$). Enfin, $p(s)$ désigne l'ensemble des objets en relation avec s pour la propriété p : $p(s) = \{o \in \mathcal{E} \,:\, \langle s, p, o\rangle \in \mathcal{K}\}$. Par exemple, $\texttt{creator}^{-1}(\texttt{Picasso})$ donne les créations de Pablo Picasso. Pour mesurer l'importance d'une entité $e \in E$ parmi un ensemble $E \subseteq \mathcal{E}$, nous voulons comptabiliser le nombre d'objets rattachés à cette entité (i.e., $|p(e)|$), mais uniquement pour une propriété p qui fait sens pour les entités E (à l'inverse des scores de la littérature qui considèrent tous les objets rattachés à l'entité par au moins une propriété). Par exemple, $\left|\texttt{creator}^{-1}(\texttt{Picasso})\right| = 1147$ correspond aux 1 147 oeuvres créées par Pablo Picasso. Le défi consiste à déterminer si une propriété p est adaptée pour être un indicateur de classement pour classer les entités de l'ensemble $E \subseteq \mathcal{E}$. En d'autres termes, lorsqu'on observe que $|p(e_1)| > |p(e_2)|$ pour deux entités $e_1 \in E$ et $e_2 \in E$, l'entité e_1 est jugée plus importante que l'entité e_2 par rapport à p. Typiquement, il parait raisonnable de classer les artistes en fonction de leur nombre d'oeuvres pour savoir lesquels ont été les plus prolifiques en utilisant la propriété $\texttt{creator}^{-1}$. A l'inverse, le nombre de pères d'un artiste avec la propriété `father` est peu intéressant car non discriminant.

De nombreux indicateurs de classement en infométrie reposent sur des quantités suivant une loi de puissance (Egghe, 2005) (e.g., le nombre de citations pour les articles ou le nombre de liens entrants pour les pages Web). L'émergence de ces distributions se justifie essentiellement par la génération continue de nouvelles entités selon un processus d'attachement préférentiel (Barabási et Albert, 1999). Un processus d'attachement préférentiel répartit les objets entre les entités en fonction de ce qu'elles ont déjà, de sorte que celles qui ont déjà beau-

4. Cet exemple comme l'ensemble des illustrations est tiré de Wikidata.

coup d'objets reçoivent plus que celles qui en ont peu. Par exemple, un nouvel article a plus de chance de citer un article bien connu et donc déjà fortement cité. Si ce processus d'attachement préférentiel nous parait aisément transposable pour certaines propriétés, il ne couvre pas toutes les propriétés pertinentes (cf. les expérimentations en section 4). Typiquement, une nouvelle peinture ne pourra plus être créée par Pablo Picasso malgré son oeuvre prolifique [5]. Même si les lois de puissances peuvent être induites par d'autres phénomènes (Newman, 2005), il nous semble difficilement justifiable de se restreindre uniquement aux distributions suivant une loi de puissance pour construire des classements – d'autant qu'en pratique, de telles distributions se distinguent difficilement des autres, comme d'une distribution log-normale par exemple (Mitzenmacher, 2004). Par contre, il nous parait essentiel de constater qu'une distribution suivant une loi de puissance (ou même une distribution log-normale) concentre une grande quantité sur peu d'entités au détriment du reste des entités. C'est même cet écart qui explique bien souvent l'attrait du classement. Par exemple, le classement des milliardaires est spectaculaire car une poignée de riches concentrent toute la richesse.

De ce fait, nous faisons l'hypothèse qu'une propriété est propice à mesurer l'impact d'entités si sa distribution des objets est inégalitairement répartie. En effet, si les objets d'une propriété sont répartis de manière relativement uniforme, l'écart réduit en nombre d'objets entre deux entités risque d'être insuffisant pour justifier que l'une soit considérée comme meilleure que l'autre. Pire, il risque d'y avoir beaucoup d'entités ex aequo même parmi les entités du haut du classement. Plus formellement, pour n entités $E = \langle e_1, e_2, \ldots, e_n \rangle$ où $|p(e_i)| \leq |p(e_{i+1})|$, nous considérons qu'une propriété p donne un indicateur de classement pertinent pour E si les deux critères suivants sont vérifiés :

C1 L'écart entre celui qui a le moins d'objets pour la propriété p (i.e., e_1) et celui qui en a le plus (i.e., e_n) est élevé : $|p(e_1)| \ll |p(e_n)|$.

C2 La distribution des objets pour la propriété p est plus inégalitaire qu'une distribution uniforme entre $|p(e_1)|$ et $|p(e_n)|$.

En d'autres termes, le critère C1 explicite un fort écart de richesse et le critère C2, une répartition plus injuste que le hasard. Par exemple, le critère C2 exclut une propriété où tout le monde serait riche sauf e_1. Pour mesurer le niveau d'inégalité d'une distribution évoqué par le critère C2, nous devons recourir à une mesure de concentration (Cowell, 2011). Nous optons pour le coefficient de Gini (1936) qui est une mesure de concentration régulièrement utilisée en économie pour estimer les inégalités de revenus, mais aussi en bibliométrie (Pratt, 1977; Rousseau, 1994) [6]. Plus précisément, le coefficient de Gini, ou indice de Gini, est une mesure statistique évaluant le niveau d'inégalité de la répartition d'une variable dans une population. Il augmente entre 0 et 1 avec le niveau d'inégalité, où 0 signifie l'égalité parfaite et 1, l'inégalité parfaite. Dans notre contexte, le coefficient de Gini d'une propriété p pour n entités $E = \langle e_1, e_2, \ldots, e_n \rangle$ où $|p(e_i)| \leq |p(e_{i+1})|$ se calcule de la manière suivante :

$$G_p(E) = \frac{2 \sum_{i=1}^{n} i \times |p(e_i)|}{n \sum_{i=1}^{n} |p(e_i)|} - \frac{n+1}{n} \quad (1)$$

5. De toute façon, en suivant le protocole de Clauset et al. (2009), le nombre d'oeuvres par artiste ne semble pas satisfaire une loi de puissance puisqu'avec les paramètres les plus favorables (i.e., l'exposant $\alpha = 2.13$ et $x_{min} = 16$), la différence est de 0.031, bien au-dessus du seuil 0.019 correspondant à la valeur critique.

6. La même méthodologie peut être suivie avec d'autres mesures d'inégalité comme l'indice d'Atkinson (Atkinson et al., 1970) ou l'index d'entropie généralisée (Shorrocks, 1980).

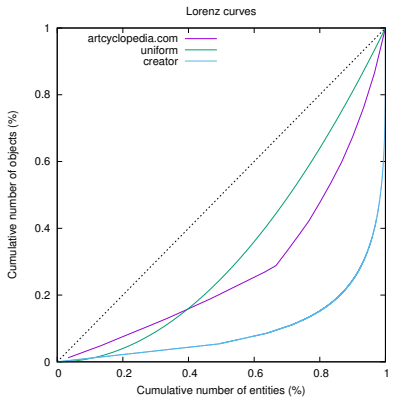

FIG. 1: Courbes de Lorenz d'une distribution uniforme (en bleu) et de la propriété $\texttt{creator}^{-1}$ (en vert) et de l'indice de `artcyclopedia.com` (en violet)

Rang	Peintre	Nbr. peintures	Rang	Peintre	Nbr. peintures
1	Kalervo Palsa	1 940	11	Mary Vaux Walcott	787
2	Edvard Munch	1 789	12	David Teniers the Younger	718
3	Anthony van Dyck	1 215	13	Pablo Picasso	705
4	John Everett	1 021	14	Rembrandt	690
5	Peter Paul Rubens	1 001	15	Paul Cézanne	656
6	George Catlin	968	16	Camille Pissarro	644
7	Claude Monet	940	17	Panos Terlemezian	631
8	Pierre-Auguste Renoir	907	18	Eliseu Visconti	590
9	Vincent van Gogh	890	19	Edwin Austin Abbey	585
10	Jean-Baptiste-Camille Corot	879	20	Emanuel Fohn	549

TAB. 2: Classement des peintres issu de la propriété $\texttt{creator}^{-1}$ restreinte aux peintures

Nous dénotons aussi par G_p le coefficient de Gini où l'ensemble d'entités E correspond à toutes les entités de $\mathcal{E}$ ayant au moins un fait pour la propriété p. Typiquement, $G_{\texttt{creator}^{-1}} = 0,805$ correspond au coefficient de Gini de toutes les entités ayant au moins une création. Le coefficient de Gini peut également se représenter graphiquement comme deux fois l'aire entre l'identité et la courbe de Lorenz. La figure 1 représente la courbe de Lorenz pour la propriété $\texttt{creator}^{-1}$ en bleu (i.e., pour k entités parmi les créateurs, l'ordonnée représente $\sum_{i=1}^{k} \left|\texttt{creator}^{-1}(e_i)\right|$ en pourcentage). Dans cette représentation, l'égalité parfaite correspond à l'identité (en tirets) et l'inégalité la plus forte à la courbe $\{(0,0),(1,0),(1,1)\}$. Finalement, la table 2 donne le classement des peintres suivant leur nombre de peintures. Bien entendu, ce classement diffère de ceux proposés par la table 1 car les peintres les plus prolifiques ne sont pas forcément les plus importants.

A ce stade, il ne fait aucun doute que la propriété $\texttt{creator}^{-1}$ engendre un classement pertinent à cause de son coefficient de Gini élevé. Mais, est-ce aussi le cas pour le classement du site `artcyclopedia.com` dont l'index a un coefficient de Gini [7] de seulement 0.442 ? Il

7. Le coefficient de Gini a été calculé à partir des indices des 30 peintres les plus importants. Si nous disposions de toutes les données, la valeur aurait probablement été supérieure.

nous faut déterminer à partir de quel niveau d'inégalité une mesure est jugée pertinente pour établir un classement. En reprenant les deux critères C1 et C2, il est aisé de démontrer que le coefficient de Gini [8] de la propriété p pour les entités E doit être supérieur à un tiers : $G_p(E) \geq 1/3$. En effet, pour une distribution uniforme (critère C2), le coefficient de Gini est égal à $\frac{|p(e_n)|-|p(e_1)|}{3(|p(e_n)|+|p(e_1)|)}$ or le critère C1 nous donne que $|p(e_n)|-|p(e_1)| \approx |p(e_n)|+|p(e_1)|$. Sur la figure 1, la courbe de Lorenz issue des critères C1 et C2 est tracée en vert. Bien sûr, l'aire de la courbe bleue est bien plus importante (car $G_{\texttt{creator}^{-1}} = 0.805 > 1/3$). Nous concluons aussi que le classement par popularité du site `artcyclopedia.com` fait sens puisque son coefficient de Gini est de 0.442 (courbe violette). Néanmoins, il est basé sur des données d'usage qui ne possèdent pas d'aussi bonnes propriétés que le Web des données comme listées dans la section précédente. En particulier, le classement issu de la propriété $\texttt{creator}^{-1}$ est indépendant de toute représentation informatique. Si un expert en peinture avait comptabilisé manuellement les toiles des différents maitres, il parviendrait à un classement similaire. A l'inverse, le classement fondé sur les données d'usage n'a de sens qu'à travers la consommation des ressources informatiques.

4 Cas d'étude sur Wikidata

Wikidata est une base de connaissances libre éditée de manière collaborative (Vrandečić et Krötzsch, 2014). Comme cette base de connaissances généraliste concerne des entités très variées (personne, lieux, événements, etc), elle est idéale pour mettre à l'épreuve la transdisciplinarité de notre approche. Après avoir détaillé le protocole expérimental, nous comparons le modèle basé sur l'inégalité au modèle de la loi de puissance. Ensuite, nous illustrons l'intérêt de l'approche en présentant les meilleurs indicateurs de classement identifiés sur l'ensemble de Wikidata, puis sur des classes et professions spécifiques. Le code source (réalisé en Java avec la librairie Jena) et le résultat de l'exécution sur Wikidata est disponible à l'adresse suivante : `https://github.com/asoulet/egc21ranking`.

Jusqu'ici, nous avons implicitement fait l'hypothèse que l'addition des objets pour une entité et une propriété données (i.e., $|p(e)|$) avait toujours du sens. Pour que cela soit vrai, la propriété doit être additive (Lenz et Shoshani, 1997). L'additivité consiste à ne pas compter deux fois le même objet ce qui peut arriver si une information est répétée dans la base de connaissances. Ces répétitions interviennent souvent lorsqu'une information est réifiée dans plusieurs contextes (par exemple, pour différentes périodes). Pour cette raison, nous nous restreignons uniquement aux propriétés directes correspondant à l'espace de noms `http://www.wikidata.org/prop/direct/` (préfixe `wdt`) et où les objets sont des URI. En juin 2020, il y avait 2482 propriétés directes dont seulement 659 avaient des URI pour objets. Ensuite, les statistiques sur Wikidata montrent qu'une grande partie des propriétés ont une arité faible à savoir que peu d'objets o_i sont reliés à une même entité e pour une propriété p. A l'inverse, les entités e_i reliées à un même objet o pour une propriété p s'avèrent régulièrement nombreux. Typiquement, la propriété `creator` renseigne le plus souvent une seule entité par création (le créateur), mais de nombreuses créations sont rattachées au même créateur. De ce fait, pour chaque propriété directe p, nous appliquons notre méthode sur la propriété inverse

8. Encore une fois, la même démarche méthodologique s'applique à d'autres mesures d'inégalité. Par exemple, pour l'indice d'Atkinson avec $\epsilon = 1$, les critères C1 et C2 impliquent un seuil de $1 - 2/e$.

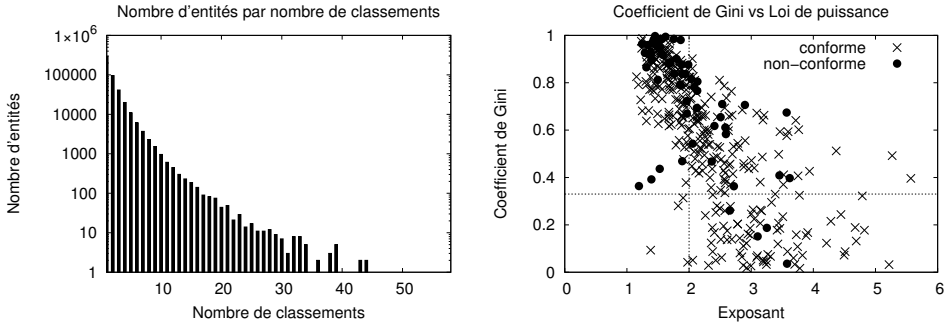

FIG. 2: Histogramme du nombre d'entités suivant leur nombre de classements et comparaison avec la loi de puissance

p^{-1} plus propice à l'émergence d'inégalités et donc, d'un indicateur de classement pertinent (comme nous l'avons fait précédemment avec `creator`$^{-1}$). Au final, nous avons calculé le coefficient de Gini pour l'inverse des 659 propriétés candidates. 386 propriétés inverses ont un coefficient supérieur à $1/3$ dont 362 classent au moins 10 entités. Ces propriétés permettent de classer 2 200 983 entités. L'histogramme de la figure 2 reporte le nombre d'entités suivant leur nombre de classements. La majorité des entités (1 724 525 sur 2 200 983) n'appartiennent qu'à un seul classement. A l'autre extrême, les Etats-Unis d'Amérique sont classés via 58 propriétés.

Notre première expérience consiste à évaluer la pertinence de nos critères C1 et C2 face au modèle de la loi de puissance. Pour chaque propriété, nous avons modélisé la distribution par la meilleure loi de puissance $cx^{-\alpha}$ (où c est une constante et α l'exposant) en utilisant le protocole de Clauset et al. (2009). Le graphique de droite sur la figure 2 trace le coefficient de Gini en fonction de l'exposant retenu pour chaque propriété (les croix sont conformes et les points non-conformes). Premièrement, les modèles utilisés en infométrie reposant sur l'attachement préférentiel (Barabási et Albert, 1999) impliquent un exposant supérieur ou égal à 2 qui ne sont pas adaptés à la majorité des propriétés suivant une loi de puissance (i.e., les croix à gauche de la droite $x = 2$). Deuxièmement, 24% des propriétés avec un coefficient de Gini supérieur à 0.80 ne sont pas conformes à une loi de puissance (cf. la concentration de points de la figure 2). Par exemple, la moitié des propriétés de la table 3 (signifiée par un astérisque) ne suivent pas une loi de puissance. La loi de puissance est contradictoire en ne considérant pas la propriété `P407` comme intéressante alors que 2 autres propriétés très similaires sont jugées intéressantes (`P1412` et `P103`). Troisièmement, de nombreuses propriétés suivant une loi de puissance ont un coefficient de Gini inférieur au seuil $1/3$ enfreignant au moins un des critères C1 ou C2. Une grande partie de ces propriétés classent des images, des cartes ou d'autres ressources informatiques qui n'existent pas dans le monde réel. Ces trois constats confortent notre choix de recourir aux critères C1 et C2 sans faire d'hypothèse de modèle paramétrique suivant une loi de puissance.

La table 3 indique les 20 propriétés ayant les coefficients de Gini les plus élevés. On constate que ces propriétés n'ont pas de cardinalité maximale ce qui est propice à de fortes inégalités. Par exemple, de forts écarts sont attendus entre le nombre d'entités anglophones et

Propriété p	Gini G_{p-1}	#entités	Propriété p	Gini G_{p-1}	#entités
*lang. of work or name (P407)	0.997	1172	*occupation (P106)	0.982	12 147
instance (P31)	0.995	65 777	country of citizenship (P27)	0.981	3 093
*writing system (P282)	0.994	447	*religion (P140)	0.981	1 245
epoch (P6259)	0.994	168	country of origin (P495)	0.981	1 252
sex or gender (P21)	0.987	113	*described by source (P1343)	0.981	28 828
color (P462)	0.986	242	*instrument (P1303)	0.98	818
*manner of death (P1196)	0.985	309	*collection (P195)	0.98	7 457
*material used painting (P186)	0.985	4 163	*subject has role (P2868)	0.979	878
lang. spoken or written (P1412)	0.984	1 374	honorific prefix (P511)	0.978	196
country (P17)	0.983	2 553	native language (P103)	0.976	784

TAB. 3: Les 20 meilleures propriétés de Wikidata pour classer les entités selon l'indice de Gini

Classe	Propriété p pour classer	Gini G_{p-1}	#classés	#instances
Human (Q5)	creator (P170)	0.805	90	718 384
Taxon (Q16521)	parent taxon (P171)	0.781	91	88 720
position (Q4164871)	position held (P39)	0.877	88	81 046
Scientific journal (Q5633421)	published in larger work (P1433)	0.913	99	30 146
Communes of France (Q484170)	destination point (P1444)	0.531	17	29 629
Surname (Q101352)	second family name (P1950)	0.657	69	29 175
Business (Q4830453)	manufacturer (P176)	0.811	57	24 398
Band (rock and pop) (Q215380)	performer (P175)	0.672	14	22 774
River (Q4022)	mouth of the watercourse (P403)	0.614	52	21 642
scholarly article (Q13442814)	described by source (P1343)	0.981	12	19 256
association football club (Q476028)	participating team (P1923)	0.787	68	17 253
Gene (Q7187)	regulates (molecular biology) (P128)	0.63	57	14 458
Road (Q34442)	connects with (P2789)	0.487	36	13 784
Immortalised cell line (Q21014462)	parent cell line (P3432)	0.86	60	13 013
Organization (Q43229)	member (P463)	0.838	10	10 977
Award (Q618779)	award received human (P166)	0.896	17	9 534
Season (sports) (Q27020041)	participant (P1344)	0.787	13	8 005
language (Q34770)	language of work or name (P407)	0.997	88	7 788
Comune (Q747074)	ancestral home (P66)	0.419	11	7 731
University (Q3918)	affiliation (P1416)	0.827	59	7 678

TAB. 4: Les 20 plus grandes classes de Wikidata parmi les entités classées avec leur classement

le nombre d'entités parlant quechua pour les propriétés relatives à la langue (i.e., P407, P1412, P103). De manière intéressante, les meilleures propriétés mesurent l'impact d'entités appartenant à des disciplines bien différentes : la linguistique, l'histoire, la géographie, la musique, la religion, etc. Ensuite, les tables 4 et 5 détaillent respectivement les 20 classes et les 20 professions les plus importantes de Wikidata [9]. Pour chaque catégorie (i.e., classe ou profession), il est indiqué la propriété inverse qui a classé le plus d'entités de cette catégorie parmi les 100 premiers (correspondant à la colonne #classés, e.g., creator est la propriété classant le plus d'humains dans les 100 premiers à savoir 90). Ainsi, la première ligne de la table 4 suggère qu'il est possible de classer les humains décrits dans Wikidata en comptabilisant leur nombre de créations. Pour des professions plus spécifiques, la table 5 propose d'autres classements qui paraissent bien raisonnables au regard de ces professions (par exemple, compter le nombre de castings pour les acteurs ou de films réalisés pour les réalisateurs). On retrouve aussi le classement des peintres suivant leur nombre de créations. A nouveau, la table 4 montre la diversité des disciplines considérées et en même temps, la spécificité disciplinaire des propriétés

9. Ces classements issus des propriétés P31 et P106 font sens car leur coefficient de Gini est élevé (cf. la table 3).

Profession	Propriété p pour classer	Gini G_{p-1}	#classés	#instances
Actor (Q33999)	cast member (P161)	0.674	50	79 386
Writer (Q36180)	screenwriter (P58)	0.577	31	55 002
Painter (Q1028181)	creator (P170)	0.805	75	45 771
Film director (Q2526255)	director (P57)	0.643	61	40 620
film actor (Q10800557)	cast member (P161)	0.674	64	38 444
Screenwriter (Q28389)	director (P57)	0.643	58	36 938
Singer (Q177220)	performer (P175)	0.672	43	29 127
television actor (Q10798782)	cast member (P161)	0.674	40	25 231
Composer (Q36834)	composer (P86)	0.736	72	24 346
university teacher (Q1622272)	author (P50)	0.859	22	24 076
Journalist (Q1930187)	presenter (P371)	0.457	21	22 884
Film producer (Q3282637)	director (P57)	0.643	50	18 325
Catholic priest (Q250867)	consecrator (P1598)	0.528	48	17 619
sport cyclist (Q2309784)	classification of race participants (P2321)	0.694	72	17 540
Architect (Q42973)	architect (P84)	0.544	42	17 268
Poet (Q49757)	author (P50)	0.859	19	16 417
stage actor (Q2259451)	cast member (P161)	0.674	33	15 919
Musician (Q639669)	composer (P86)	0.736	26	15 130
Historian (Q201788)	editor (P98)	0.378	15	10 814
sculptor (Q1281618)	creator (P86)	0.804	21	9 784

TAB. 5: Les 20 professions les plus usuelles de Wikidata avec leur classement

choisies pour mesurer l'impact (e.g., les gènes classés suivant leur nombre de molécules les régulant). De manière intéressante, notre méthode découvre automatiquement des indicateurs de classement utilisés en scientométrie : nombre d'articles publiés par un journal scientifique ou nombre de références à un article, nombre d'articles publiés par un universitaire, etc. D'autres indicateurs feraient aussi sens comme la propriété `parent taxon` en glottométrie ou la propriété `manufacturer` en économétrie.

5 Discussion et perspectives

La méthodologie présentée dans cet article montre comment découvrir automatiquement les quantités inégalitaires dans une discipline conduisant à des indicateurs de classement grâce à un usage original des mesures de concentrations. Le cas d'étude a validé le bien-fondé de cette méthodologie. Même s'il s'agit d'un premier pas, nous pensons que les quantités trouvées sont fondamentales pour élaborer des indicateurs plus subtils. Pour tenir compte de la qualité des objets comptés (et pas seulement des aspects quantitatifs), il est possible de combiner les classements. Typiquement, le h-index consiste à compter le nombre de publications bien classées en nombre de citations (Hirsch, 2005). De la même manière, une entité e pourrait avoir un (p, q)-index h si l'entité dispose de h objets pour p ayant chacun plus de h objets pour q et tous les autres objets ont au plus h objets pour q. Par exemple, la table 6 classe les peintres suivant le $(\texttt{creator}^{-1}, \texttt{schema:about}^{-1})$-index h. Même s'il diffère des classements de la table 1, l'impact mesuré élimine les peintres moins unanimement reconnus.

Au-delà des indicateurs de classement, la valeur stratégique du Web des données mérite à nos yeux un plus grand intérêt. La principale leçon de ce travail est la nécessité de modèles transdisciplinaires comme celui fondé sur les critères C1 et C2 qui ne s'appuie pas sur des hypothèses spécifiques à un domaine comme par exemple la loi de Lotka en science de l'information. Même si notre modèle décrit moins précisément les phénomènes car s'appuyant

Rg	Peintre	h	Rg	Peintre	h	Rg	Peintre	h
1	Vincent van Gogh	17	11	Hieronymus Bosch	11	21	Jacques-Louis David	10
2	Leonardo da Vinci	16	12	Jan van Eyck	11	22	Paul Gauguin	10
3	Johannes Vermeer	16	13	Pierre-Auguste Renoir	11	23	Caspar David Friedrich	9
4	Raphael	15	14	J.-A.-D. Ingres	11	24	Rogier van der Weyden	9
5	Pieter Bruegel the Elder	15	15	Gustave Courbet	10	25	Pablo Picasso	9
6	Caravaggio	15	16	Joseph Wright of Derby	10	26	Claude Monet	9
7	Rembrandt	12	17	El Greco	10	27	W.-A. Bouguereau	9
8	Titian	12	18	Sandro Botticelli	10	28	Albrecht Dürer	9
9	Édouard Manet	12	19	Francisco Goya	10	29	Gustav Klimt	8
10	Diego Velázquez	11	20	Salvador Dalí	10	30	Jan Matejko	8

TAB. 6: Classement des peintres inspiré de l'index Hirsch

sur moins d'expertise, son application étend la science de la mesure à un large ensemble de domaines de la connaissance, bien au-delà de l'infométrie ou de la scientométrie.

Références

Atkinson, A. B. et al. (1970). On the measurement of inequality. *Journal of economic theory 2*(3), 244–263.

Barabási, A.-L. et R. Albert (1999). Emergence of scaling in random networks. *Science 286*(5439), 509–512.

Behn, R. D. (2003). Why measure performance ? Different purposes require different measures. *Public administration review 63*(5), 586–606.

Berners-Lee, T., J. Hendler, et O. Lassila (2001). The semantic web. *Scientific american 284*(5), 34–43.

Clauset, A., C. R. Shalizi, et M. E. Newman (2009). Power-law distributions in empirical data. *SIAM review 51*(4), 661–703.

Cowell, F. (2011). *Measuring inequality*. Oxford University Press.

Ding, L., T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, et J. Sachs (2004). Swoogle : a search and metadata engine for the semantic web. In *International Conference on Information and Knowledge Management*, pp. 652–659.

Egghe, L. (2005). *Power laws in the information production process : Lotkaian informetrics*. Emerald.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science 178*, 471–479.

Gini, C. (1936). On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series 208*, 73–79.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences 102*(46), 16569–16572.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM) 46*(5), 604–632.

Lenz, H.-J. et A. Shoshani (1997). Summarizability in OLAP and statistical data bases. In *International Conference on Scientific and Statistical Database Management*, pp. 132–143.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet mathematics 1*(2), 226–251.

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton University Press.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics 46*(5), 323–351.

Page, L., S. Brin, R. Motwani, et T. Winograd (1999). The pagerank citation ranking : Bringing order to the web. Technical report, Stanford InfoLab.

Pinski, G. et F. Narin (1976). Citation influence for journal aggregates of scientific publications : Theory, with application to the literature of physics. *Information processing & management 12*(5), 297–312.

Pratt, A. D. (1977). A measure of class concentration in bibliometrics. *Journal of the American Society for Information Science 28*(5), 285–292.

Razniewski, S., F. Suchanek, et W. Nutt (2016). But what do we actually know ? In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, pp. 40–44.

Rousseau, R. (1994). Similarities between informetrics and econometrics. *Scientometrics 30*(2-3), 385–387.

Shorrocks, A. F. (1980). The class of additively decomposable inequality measures. *Econometrica : Journal of the Econometric Society*, 613–625.

Soulet, A., A. Giacometti, B. Markhoff, et F. M. Suchanek (2018). Representativeness of knowledge bases with the generalized benford's law. In *ISWC*, pp. 374–390. Springer.

Stuart, D. (2014). *Web metrics for library and information professionals*. Facet Publishing.

Thelwall, M., L. Vaughan, et L. Björneborn (2005). Webometrics. *Annual review of information science and technology 39*(1), 81–135.

Vrandečić, D. et M. Krötzsch (2014). Wikidata : a free collaborative knowledgebase. *Communications of the ACM 57*(10), 78–85.

Weinberg, A. M. (2000). Criteria for scientific choice (minerva, i (2),(1962), 158–171). *Minerva 38*(3), 253–266.

Zaveri, A., A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, et S. Auer (2016). Quality assessment for linked data : A survey. *Semantic Web 7*(1), 63–93.

Summary

Analyzing the impact of entities within their field is fundamental to understand it. To this end, it is essential to have fine numerical indicators retranscribing the specificities of the field. This paper proposes a transdisciplinary approach to automatically discover ranking scores having nevertheless an intradisciplinary semantics. For this purpose, our approach is based on the knowledge bases of the Web of data, not only to facilitate the operational computation of the indicators but also to take advantage of their transparency and their semantic richness. The simple but central hypothesis of this work is that each unequal distribution of a quantity generates a relevant ranking score. To this end, we use the Gini coefficient to identify in Wikidata the properties producing significant ranking scores.