

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
29 December 2022 (29.12.2022)



(10) International Publication Number
WO 2022/268808 A1

- (51) **International Patent Classification:**
G06F 16/35 (2019.01) *G06V 10/82* (2022.01)
G06F 16/951 (2019.01)
- (21) **International Application Number:**
PCT/EP2022/066874
- (22) **International Filing Date:**
21 June 2022 (21.06.2022)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
17/359,082 25 June 2021 (25.06.2021) US
- (71) **Applicant: META CLUSTER LT, UAB** [LT/LT]; A.
Gostauto g. 40A, 03163 Vilnius (LT).
- (72) **Inventors: JURAVICIUS, Martynas;** Naugarduko Street
62-85, 03203 Vilnius (LT). **KUKSTA, Andrius;** Raugyk-
los Street 17-23, 01140 Vilnius (LT).
- (74) **Agent: MEISSNER BOLTE PATENTANWÄLTE**
RECHTSANWÄLTE PARTNERSCHAFT MBB; Mr.
Florian Meyer, Widenmayerstr. 47, 80358 Munich (DE).
- (81) **Designated States** (*unless otherwise indicated, for every
kind of national protection available*): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ,
CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO,
DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN,
HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH,
KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA,
MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI,
NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU,
RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM,

(54) **Title:** ADVANCED RESPONSE PROCESSING IN WEB DATA COLLECTION

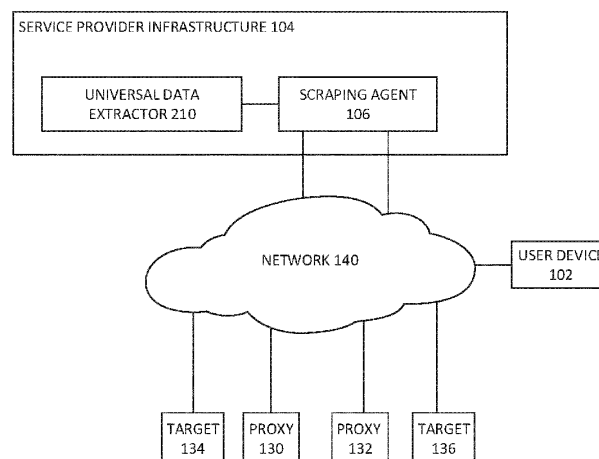


FIGURE 1

(57) **Abstract:** Advanced response processing in web data collection discloses processor-implemented apparatuses, methods, and systems of processing unstructured raw HTML responses collected in the context of a data collection service, the method comprising, in one embodiment, receiving raw unstructured HTML documents and extracting text data with associated meta information that may comprise style and formatting information. In some embodiments data field tags and values may be assigned to the text blocks extracted, classifying the data based on the processing of Machine Learning algorithms. Additionally, blocks of extracted data may be grouped and re-grouped together and presented as a single data point. In another embodiment the system may aggregate and present the text data with the associated meta information in a structured format. In certain embodiments the Machine Learning model may be a model trained on a pre-created training data set labeled manually or in an automatic fashion.

[Continued on next page]

TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report (Art. 21(3))*
- *with amended claims (Art. 19(1))*
- *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*