



- (51) **International Patent Classification:**
G06F 16/958 (2019.01) *G06F 16/951* (2019.01)
- (21) **International Application Number:**
PCT/EP2018/085983
- (22) **International Filing Date:**
19 December 2018 (19.12.2018)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
17209553.1 21 December 2017 (21.12.2017) EP
- (71) **Applicant:** URBAN SOFTWARE INSTITUTE GMBH
[DE/DE]; Zwickauer Str. 223a, 09116 Chemnitz (DE).
- (72) **Inventors:** BITTNER, Christopher; Hohenzollerndamm
25, 10717 Berlin (DE). MÜLLER, Sebastian; Sächsische
Straße 23, 10717 Berlin (DE).
- (74) **Agent:** BITTNER, Peter et al.; Seegarten 24, 69190 Wall-
dorf (DE).

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, SM, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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))

- (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, IR, IS, JO, JP, KE, KG, KH, KN, KP,

- (54) **Title:** COMPUTER SYSTEM AND METHOD FOR EXTRACTING DYNAMIC CONTENT FROM WEBSITES

(57) **Abstract:** Computer system (100), computer-implemented method and computer program product are provided for extracting dynamic content data (221) from a website (220) in a machine-readable format. The system has an interface (110) to access configuration data (250) reflecting the structure of the website (220). The configuration data includes at least a website specific scraping script and one or more website specific XPath statements. Further, the interface receives a data retrieval request (210) specifying the website (220) and corresponding dynamic content data (221) to be retrieved. A scraper module (120) provides the scraping script (2050) for execution wherein the scraping script is configured to perform one or more parameterized navigation steps on the website (220) to access the dynamic content data (221). A script module (140) triggers execution of the scraping script and receives HTML/XML data associated with the dynamic content data from the website (220) in response to the scraping script execution. An XPath extraction module (150) extracts machine-readable content data (222) from the HTML/XML data wherein the XPath extraction module is pre-configured with the website specific XPath statements in accordance with the structure of the website (220).

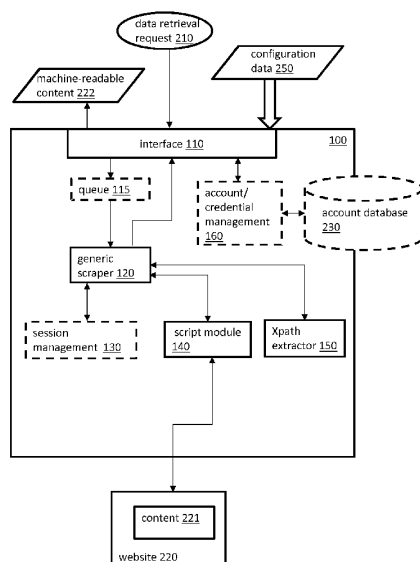


FIG. 1